Introduction to Statistics: A Simulation-First Approach

An Open Educational Resource

Jeff Baggett

2026-10-20

Table of contents

Welcome **1**
Why simulation first? . . . 1
How to use this book . . . 2
About this book . . . 2
License . . . 2

Preface **3**
For instructors . . . 3
StatLens integration . . . 3
Accessibility . . . 3
Acknowledgments . . . 4

I Unit 1: Explore **5**

1 Introduction to Data **7**
1.1 Case study: Using stents to prevent strokes . . . 7
1.2 Data basics . . . 9
1.3 Types of variables . . . 11
1.4 Relationships between variables . . . 14
1.5 Explanatory and response variables . . . 16
1.6 Observational studies and experiments . . . 17
1.7 Chapter review . . . 19
1.8 Exercises . . . 20

2 Data Collection and Study Design **34**
2.1 Populations and samples . . . 34
2.2 Sampling from a population . . . 36
2.3 Random sampling methods . . . 39
2.4 Experiments . . . 43
2.5 Observational studies . . . 46
2.6 Confounding variables . . . 46
2.7 Random assignment and causation . . . 47
2.8 Scope of inference . . . 48
2.9 Chapter review . . . 49
2.10 Exercises . . . 51

3 Exploring Categorical Data **63**
3.1 Visualizing one categorical variable . . . 63
3.2 Two categorical variables . . . 66
3.3 Effective data visualization . . . 70
3.4 Chapter review . . . 74

3.5 Exercises . . . 74

4 Exploring Quantitative Data **82**
4.1 Visualizing quantitative data . . . 82
4.2 Measures of center . . . 89
4.3 Measures of spread . . . 92
4.4 Outliers and robust statistics . . . 95
4.5 Box plots . . . 96
4.6 Describing a distribution . . . 98
4.7 The empirical rule and Chebyshev's theorem . . . 98
4.8 Chapter review . . . 100
4.9 Exercises . . . 101

5 Relationships Between Variables **116**
5.1 Comparing groups . . . 116
5.2 Scatterplots . . . 119
5.3 Association vs. causation . . . 122
5.4 Chapter review . . . 123
5.5 Exercises . . . 123

Explore — Unit Review **129**
The big picture . . . 129
Which display and which summary? . . . 129
A quick decision guide . . . 130
Key ideas to carry forward . . . 130
Common pitfalls . . . 130
Exercises . . . 130

Tech Tutorial: Explore **132**
The data . . . 132
What you'll do . . . 132
1. Get to know the data . . . 133
2. Summarize one categorical variable . . . 133
3. Summarize one numerical variable . . . 134
4. Compare a numerical variable across groups . . . 136
5. Explore a relationship between two numerical variables . . . 137
Check yourself . . . 138

Project: Exploring a Distribution **139**
The question . . . 139
The data . . . 139
Plan it first . . . 139
Do the analysis . . . 139
Your deliverable . . . 141

II Unit 2: Simulate **143**

6 Sampling Variability **145**
6.1 Parameters vs. statistics . . . 145
6.2 Sampling distributions . . . 147
6.3 Variability of sample statistics . . . 148
6.4 Standard error . . . 150
6.5 Looking ahead . . . 152
6.6 Chapter review . . . 152

6.7 Exercises . . . 153

7 Randomization Tests **163**
7.1 The hypothesis testing framework . . . 163
7.2 Null and alternative hypotheses . . . 164
7.3 Case study: sex discrimination . . . 164
7.4 Building a randomization distribution . . . 166
7.5 Case study: opportunity cost . . . 168
7.6 The same idea, for means . . . 170
7.7 p-values . . . 172
7.8 Statistical discernibility . . . 172
7.9 Two-sided tests . . . 173
7.10 The randomization test procedure: a summary . . . 174
7.11 Chapter review . . . 175
7.12 Exercises . . . 176

8 Decision Errors **184**
8.1 Type I error . . . 184
8.2 Type II error: missing a real effect . . . 185
8.3 The decision table . . . 185
8.4 Error rates: α and β . . . 186
8.5 The tradeoff between α and β . . . 186
8.6 Choosing a discernibility level . . . 187
8.7 Looking ahead: the probability of a Type II error . . . 188
8.8 Chapter review . . . 188
8.9 Exercises . . . 189

9 Bootstrap Confidence Intervals **193**
9.1 The big idea: resampling . . . 193
9.2 Bootstrapping a sample mean . . . 193
9.3 Interpreting a confidence interval . . . 196
9.4 The percentile method . . . 198
9.5 The standard error method . . . 198
9.6 Medical consultant case study . . . 199
9.7 Tappers and listeners case study . . . 200
9.8 What affects the width of a CI? . . . 201
9.9 When does bootstrapping work? . . . 202
9.10 Bootstrap CI for other statistics . . . 203
9.11 Chapter review . . . 203
9.12 Exercises . . . 204

10 Sampling Distributions via Simulation **212**
10.1 From simulation to theory . . . 212
10.2 The Central Limit Theorem . . . 213
10.3 Comparing randomization, bootstrap, and mathematical approaches . . . 214
10.4 When is the normal approximation good enough? . . . 215
10.5 The normal distribution: a preview . . . 217
10.6 Recap: connecting the three approaches . . . 218
10.7 Chapter review . . . 219
10.8 Exercises . . . 220

Simulate — Unit Review **224**
The big picture . . . 224
Two engines, one idea: randomization vs. bootstrap . . . 224

Which simulation do I run? . . . 225
Key ideas to carry forward . . . 225
Common pitfalls . . . 225
Exercises . . . 225

Tech Tutorial: Simulate **228**
What you'll do . . . 228
1. Randomization test for a difference in means . . . 228
2. Bootstrap confidence interval for a mean . . . 230
Check yourself . . . 231

III Unit 3: Compute 232

11 Normal Approximation **234**
11.1 The normal distribution . . . 234
11.2 The 68-95-99.7 rule . . . 236
11.3 Z-scores and standardizing . . . 237
11.4 Finding probabilities from the normal distribution . . . 238
11.5 The normal approximation to sampling distributions . . . 243
11.6 When does the normal model apply – and when doesn't it? . . . 245
11.7 Chapter review . . . 246
11.8 Exercises . . . 246

12 Inference for Proportions **255**
12.1 Case study: is the complication rate lower? . . . 255
12.2 Building the null distribution by simulation . . . 256
12.3 Normal approximation for proportions . . . 258
12.4 Confidence interval for a single proportion . . . 259
12.5 Hypothesis test for a single proportion . . . 261
12.6 Difference of two proportions . . . 263
12.7 Comparing simulation-based and normal-based methods . . . 265
12.8 Explore with StatLens . . . 266
12.9 Chapter review . . . 266
12.10 Exercises . . . 267

13 Confidence Intervals for Means **292**
13.1 The t-distribution . . . 292
13.2 One-sample t confidence interval . . . 294
13.3 Two-sample t confidence interval . . . 296
13.4 Paired data confidence interval . . . 297
13.5 Connecting t-based and bootstrap confidence intervals . . . 299
13.6 Chapter review . . . 300
13.7 Exercises . . . 300

14 Hypothesis Tests for Means **313**
14.1 Testing a single mean by simulation . . . 313
14.2 The t-test framework . . . 314
14.3 One-sample t-test . . . 315
14.4 Two-sample t-test . . . 317
14.5 Paired t-test . . . 319
14.6 Connecting simulation-based and t-based tests . . . 321
14.7 Practical vs. statistical discernibility . . . 322
14.8 Chapter review . . . 323
14.9 Exercises . . . 323

Compute — Unit Review **341**
The big picture . . . 341
Which z- or t-procedure? . . . 341
Simulation or formula? . . . 342
Key ideas to carry forward . . . 342
Common pitfalls . . . 342
Exercises . . . 342

Tech Tutorial: Compute **344**
What you'll do . . . 344
1. Two-sample t-test for a difference in means . . . 344
2. Confidence interval for one proportion . . . 345
Check yourself . . . 345

Project: Inference for Proportions **347**
The question . . . 347
The data . . . 347
Plan it first . . . 347
Do the analysis . . . 348
Your deliverable . . . 349

Project: Comparing Two Means **351**
The question . . . 351
The data . . . 351
Plan it first . . . 351
Do the analysis . . . 351
Your deliverable . . . 353

IV Unit 4: Apply 355

15 Chi-Square Goodness of Fit **357**
15.1 Observed versus expected counts . . . 357
15.2 The chi-square statistic . . . 358
15.3 Simulation-based goodness of fit test . . . 359
15.4 The chi-square distribution . . . 360
15.5 Finding a p-value . . . 362
15.6 Chi-square goodness of fit test . . . 363
15.7 Evaluating goodness of fit for a distribution . . . 363
15.8 Chapter review . . . 364
15.9 Exercises . . . 364

16 Chi-Square Test of Independence **368**
16.1 Randomization test of independence . . . 368
16.2 Mathematical model for test of independence . . . 371
16.3 Comparing the simulation and chi-square approaches . . . 374
16.4 Chapter review . . . 375
16.5 Exercises . . . 375

17 Analysis of Variance **389**
17.1 Case study: MLB batting . . . 391
17.2 Randomization test for comparing many means . . . 392
17.3 Mathematical model for ANOVA . . . 396
17.4 Conditions for ANOVA . . . 398
17.5 Chapter review . . . 399

17.6 Exercises . . . 400

18 Multiple Comparisons **411**
18.1 The multiple testing problem . . . 411
18.2 The Bonferroni correction . . . 412
18.3 Tukey's Honest Significant Difference . . . 413
18.4 Interpreting post-hoc results . . . 414
18.5 Bonferroni vs. Tukey: When to use which . . . 416
18.6 Chapter review . . . 416
18.7 Exercises . . . 417

19 Correlation **420**
19.1 Scatterplots . . . 420
19.2 The correlation coefficient . . . 422
19.3 Correlation and causation . . . 425
19.4 Sampling variability of correlation . . . 426
19.5 Chapter review . . . 426
19.6 Exercises . . . 427

20 Linear Regression **437**
20.1 Fitting a line to data . . . 437
20.2 Residuals . . . 439
20.3 Least squares regression . . . 440
20.4 R-squared . . . 442
20.5 Inference for the slope . . . 442
20.6 Conditions for linear regression . . . 445
20.7 Chapter review . . . 446
20.8 Exercises . . . 447

21 Prediction and Model Fit **465**
21.1 Making predictions . . . 465
21.2 Extrapolation . . . 466
21.3 Residual diagnostics . . . 466
21.4 Leverage and influential points . . . 468
21.5 Confidence and prediction intervals . . . 469
21.6 Chapter review . . . 471
21.7 Exercises . . . 472

22 ANOVA for Regression **483**
22.1 The two-question view of regression . . . 483
22.2 Partitioning the total sum of squares . . . 483
22.3 The ANOVA table for regression . . . 484
22.4 The F-statistic and its distribution . . . 485
22.5 Worked example: possum head length on total length . . . 485
22.6 $F = t^2$ in simple regression . . . 486
22.7 Bridge to multiple regression . . . 486
22.8 Chapter review . . . 487
22.9 Exercises . . . 487

23 Multiple Regression **490**
23.1 From one predictor to many . . . 490
23.2 The multiple regression model . . . 491
23.3 Reading multiple-regression output: the loans case study . . . 491
23.4 Interpreting coefficients "holding others constant" . . . 492
23.5 Testing individual coefficients . . . 492

23.6 ANOVA for multiple regression . . . 493
23.7 Multicollinearity: the coin-dish story . . . 493
23.8 Categorical predictors: a brief note . . . 495
23.9 Diagnostics for multiple regression . . . 495
23.10 Chapter review . . . 496
23.11 Exercises . . . 496

Apply — Unit Review **500**
The big picture . . . 500
Which method? . . . 500
A quick decision guide . . . 501
Key ideas to carry forward . . . 501
Common pitfalls . . . 501
Exercises . . . 501

Tech Tutorial: Apply **503**
What you'll do . . . 503
1. Chi-square test of independence . . . 503
2. One-way ANOVA . . . 504
3. Simple linear regression . . . 504
Check yourself . . . 505

Project: Comparing Several Means (ANOVA) **506**
The question . . . 506
The data . . . 506
Plan it first . . . 506
Do the analysis . . . 506
Your deliverable . . . 508

Project: Association Between Two Categorical Variables **510**
The question . . . 510
The data . . . 510
Plan it first . . . 510
Do the analysis . . . 510
Your deliverable . . . 512

Project: Association Between Two Categorical Variables (AI version) **514**
The question . . . 514
The data . . . 514
Plan it first . . . 514
Do the analysis . . . 515
Your deliverable . . . 516

Project: Fitting and Testing a Regression Line **518**
The question . . . 518
The data . . . 518
Plan it first . . . 518
Do the analysis . . . 518
Your deliverable . . . 520

V Unit 5: Foundations 522

24 Probability Rules **524**
24.1 Defining probability . . . 524
24.2 Sample spaces and events . . . 526

24.3 The complement rule . . . 527
24.4 The addition rule . . . 528
24.5 Conditional probability . . . 530
24.6 The multiplication rule . . . 531
24.7 Independence . . . 533
24.8 Bayes' theorem (optional) . . . 533
24.9 Chapter review . . . 535
24.10 Exercises . . . 535

25 Random Variables **543**
25.1 Random variables . . . 543
25.2 Discrete random variables . . . 543
25.3 Cumulative distribution functions . . . 545
25.4 Continuous random variables . . . 546
25.5 Chapter review . . . 548
25.6 Exercises . . . 549

26 Expected Value and Variance **551**
26.1 Expected value of a random variable . . . 551
26.2 Variance of a random variable . . . 553
26.3 Properties of expected value . . . 553
26.4 Properties of variance . . . 554
26.5 Linear combinations of random variables . . . 555
26.6 Chapter review . . . 556
26.7 Exercises . . . 557

27 Binomial Distribution **559**
27.1 The binomial setting . . . 559
27.2 Computing binomial probabilities . . . 560
27.3 Mean and standard deviation . . . 561
27.4 Normal approximation to the binomial . . . 562
27.5 Chapter review . . . 564
27.6 Exercises . . . 564

28 More About the Normal Distribution **569**
28.1 Checking whether data are normal . . . 569
28.2 When to use — and not use — the normal model . . . 571
28.3 Chapter review . . . 571
28.4 Exercises . . . 572

29 Type II Error and Power **575**
29.1 Recap: the two error types . . . 575
29.2 Visualizing Type II error . . . 575
29.3 Factors that affect the Type II error rate . . . 577
29.4 Practical vs. statistical discernibility revisited . . . 579
29.5 Chapter review . . . 580
29.6 Exercises . . . 580

30 Statistical Power **582**
30.1 What is power? . . . 582
30.2 Why power matters . . . 583
30.3 Factors affecting power . . . 583
30.4 Simulation-based power estimation . . . 585
30.5 Power curves . . . 586
30.6 Typical power targets . . . 588

30.7 Computing power for a two-sample test . . . 589
30.8 Chapter review . . . 590
30.9 Exercises . . . 590

31 Sample Size Planning **592**
31.1 Why plan sample size? . . . 592
31.2 Sample size for estimating a mean . . . 593
31.3 Sample size for estimating a proportion . . . 594
31.4 Sample size for comparing two groups . . . 596
31.5 Practical considerations . . . 597
31.6 Summary of sample size formulas . . . 599
31.7 Chapter review . . . 599
31.8 Exercises . . . 599

Foundations — Unit Review **601**
The big picture . . . 601
Which distribution models this? . . . 601
What makes a study powerful? . . . 601
Key ideas to carry forward . . . 602
Common pitfalls . . . 602
Exercises . . . 602

Tech Tutorial: Foundations **604**
What you'll do . . . 604
1. Normal distribution . . . 604
2. Binomial distribution . . . 605
Check yourself . . . 605

Appendices 607

A Technology Setup **607**
A.1 Option 1 — Jamovi (point-and-click) . . . 607
A.2 Option 2 — R (code) . . . 607
A.3 Either way: StatLens for the ideas . . . 608

B Datasets **609**
B.1 From the `openintro` R package . . . 609
B.2 From other R packages . . . 610
B.3 UWL-local datasets . . . 611
B.4 Exercise datasets bundled with this book . . . 611
B.5 Using the datasets . . . 612

C StatLens Quick Reference **613**
C.1 Keyboard shortcuts . . . 613
C.2 URL parameters . . . 613

D Statistical Tables **615**
D.1 Common critical values — cheat sheet . . . 615
D.2 Standard normal (z) . . . 615
D.3 Student's t . . . 618
D.4 Chi-square (χ^2) . . . 618
D.5 F-distribution . . . 618
D.6 Alternatives . . . 618

E Exercise Solutions **622**
E.1 Chapter 1 — Introduction to Data . . . 623
E.2 Chapter 2 — Data Collection and Study Design . . . 623
E.3 Chapter 3 — Exploring Categorical Data . . . 625
E.4 Chapter 4 — Exploring Quantitative Data . . . 625
E.5 Chapter 5 — Relationships Between Variables . . . 627
E.6 Chapter 6 — Sampling Variability . . . 627
E.7 Chapter 7 — Randomization Tests . . . 628
E.8 Chapter 8 — Decision Errors . . . 629
E.9 Chapter 9 — Bootstrap Confidence Intervals . . . 630
E.10 Chapter 10 — Sampling Distributions via Simulation . . . 632
E.11 Chapter 11 — Normal Approximation . . . 633
E.12 Chapter 12 — Inference for Proportions . . . 634
E.13 Chapter 13 — Confidence Intervals for Means . . . 637
E.14 Chapter 14 — Hypothesis Tests for Means . . . 639
E.15 Chapter 15 — Chi-Square Goodness of Fit . . . 641
E.16 Chapter 16 — Chi-Square Test of Independence . . . 642
E.17 Chapter 17 — Analysis of Variance . . . 643
E.18 Chapter 18 — Multiple Comparisons . . . 645
E.19 Chapter 19 — Correlation . . . 645
E.20 Chapter 20 — Linear Regression . . . 647
E.21 Chapter 21 — Prediction and Model Fit . . . 649
E.22 Chapter 22 — ANOVA for Regression . . . 650
E.23 Chapter 23 — Multiple Regression . . . 651
E.24 Chapter 24 — Probability Rules . . . 652
E.25 Chapter 25 — Random Variables . . . 654
E.26 Chapter 26 — Expected Value and Variance . . . 654
E.27 Chapter 27 — Binomial Distribution . . . 655
E.28 Chapter 28 — More About the Normal Distribution . . . 656
E.29 Chapter 29 — Type II Error and Power . . . 657
E.30 Chapter 30 — Statistical Power . . . 658
E.31 Chapter 31 — Sample Size Planning . . . 658

Welcome

This is an open educational resource (OER) for introductory statistics, built on a **simulation-first** pedagogy. Students develop statistical reasoning through hands-on simulation before encountering formal mathematical theory.

> ▯ Help improve this book — leave feedback in the margin
>
> This textbook is a **work in progress**, and your feedback makes it better. You can comment **right on the page** — nothing to install.
>
> 1. Click the **< tab** on the right edge of any page to open the annotation sidebar (powered by a free tool called Hypothesis). The first time, create a free account.
> 2. **Select any text** in the book, click **Annotate**, and type your comment — it stays pinned to that exact sentence.
> 3. **Reviewers:** if you were invited to the **145 Textbook Review** group, pick it at the top of the sidebar (instead of "Public") so your notes reach the author.
>
> Anything helps — unclear explanations, notation you'd change, examples that fall flat, or plain typos.

Why simulation first?

Traditional introductory statistics courses begin with probability rules and distribution theory, then use those tools to build confidence intervals and hypothesis tests. Students memorize formulas without understanding *why* they work.

This textbook reverses that order:

1. **Explore data** — See patterns, describe distributions, visualize relationships.
2. **Simulate inference** — Use resampling (bootstrap, permutation) to understand *what inference is* and *why it works*.
3. **Approximate with theory** — Learn that the normal and t-distributions are mathematical shortcuts for what simulation does directly.
4. **Formalize probability** — Study probability rules and random variables *after* students already have deep intuition about randomness.
5. **Model relationships** — Apply all inference tools to regression and categorical data.

By the time students encounter the Central Limit Theorem, they have already *built* sampling distributions hundreds of times. The theorem becomes an elegant confirmation of what they've seen, not an abstract declaration to accept on faith.

How to use this book

Throughout this text, **StatLens** activities invite you to explore statistical concepts interactively. Look for boxes like this one. Each activity links to a browser-based simulation tool — no software installation required.

Launch StatLens to explore all available tools.

Guided Practice exercises appear throughout each chapter with solutions in footnotes — try them before peeking.

Worked Examples walk through complete analyses step by step.

Definitions highlight key statistical vocabulary.

About this book

This textbook remixes material from two outstanding open-source statistics textbooks:

- Introduction to Modern Statistics (2nd edition) by Mine Çetinkaya-Rundel and Johanna Hardin
- OpenIntro Statistics by David Diez, Mine Çetinkaya-Rundel, and Christopher Barr

Both are licensed under Creative Commons BY-SA 3.0, and this remix is shared under the same license, as OpenIntro asks of derivative works.

Interactive simulations are powered by StatLens, an open-source statistics visualization toolkit developed at the University of Wisconsin-La Crosse.

License

This work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License.

You are free to share and adapt this material for any purpose, even commercially, as long as you give appropriate credit and distribute your contributions under the same license.

Preface

For instructors

This textbook is designed around the **simulation-based inference** (SBI) approach popularized by Lock, Lock, Lock, Lock, and Lock in *Statistics: Unlocking the Power of Data*. The pedagogical sequence is:

EDA → Simulation inference → Theoretical inference → Probability → Modeling

The book contains more material than most one-semester courses will cover. The 28 chapters are organized into six parts, and instructors should select the chapters that fit their curriculum. A typical 15-week course might cover:

- **Part I: Explore** (Chapters 1–5): Weeks 1–3
- **Part II: Simulate** (Chapters 6–9): Weeks 4–6
- **Part III: Calculate** (Chapters 10–13): Weeks 7–9
- **Part IV: Relationships** (Chapters 14–20): Weeks 10–13
- **Part V: Foundations** (Chapters 21–25): Select chapters as needed
- **Part VI: Design** (Chapters 26–28): As time permits

Parts V (probability foundations) and VI (power/sample size) are included for courses that need them but can be omitted without loss of continuity.

StatLens integration

StatLens is a free, browser-based statistics toolkit that supports the simulation-first approach. The text features StatLens throughout but is written to be self-contained — instructors who prefer StatKey, R's `infer` package, or other tools can substitute freely.

StatLens activities in this text use URL parameters for reproducibility. For example:

`https://learnlens.org/statlens/simulate/bootstrap-mean/?seed=42&n=30`

The `seed` parameter ensures every student who uses the same link gets identical results, enabling graded simulation-based assignments.

Accessibility

This textbook is designed to meet WCAG 2.1 AA accessibility standards:

- **Text**: Atkinson Hyperlegible font, designed for readers with low vision
- **Math**: MathJax rendering with screen reader support
- **Color**: All color choices meet 4.5:1 contrast ratio; information is never conveyed by color alone

- **Structure**: Semantic HTML with proper heading hierarchy
- **Images**: All figures include descriptive alt text
- **Navigation**: Full keyboard accessibility with visible focus indicators

If you encounter accessibility barriers, please file an issue on the book's GitHub repository.

Acknowledgments

This book remixes and augments material from two open-source textbooks published by the OpenIntro project: *Introduction to Modern Statistics* (IMS) by Mine Çetinkaya-Rundel and Johanna Hardin, and *OpenIntro Statistics* by David Diez, Mine Çetinkaya-Rundel, and Christopher Barr. Their commitment to open education made this work possible.

The simulation-first pedagogy is inspired by the Lock family's pioneering work in *Statistics: Unlocking the Power of Data*.

StatLens was developed at the University of Wisconsin-La Crosse with support from the Department of Mathematics and Statistics.

PART I

UNIT 1: EXPLORE

Before we can draw conclusions from data, we need to understand what data *is*, where it comes from, and what it looks like. We develop the vocabulary and visual tools for exploring data — identifying variable types, visualizing distributions, and describing relationships. These skills form the foundation for everything that follows.

Chapter 1

Introduction to Data

Scientists seek to answer questions using rigorous methods and careful observations. These observations — collected from the likes of field notes, surveys, and experiments — form the backbone of a statistical investigation and are called **data**. Statistics is the study of how best to collect, analyze, and draw conclusions from data. In this first chapter, we focus on both the properties of data and on the collection of data. We will introduce the building blocks of statistical thinking: observations, variables, data frames, and the types of questions that statistics can help us answer.

1.1 Case study: Using stents to prevent strokes

In this section we introduce a classic challenge in statistics: evaluating the efficacy of a medical treatment. The terms and ideas introduced here will be revisited throughout the course. For now, the goal is simply to get a sense of the role statistics can play in practice.

A stent is a small mesh tube that is placed inside a narrow or weak artery to assist in patient recovery after cardiac events and reduce the risk of an additional heart attack or death. Many doctors have hoped that there would be similar benefits for patients at risk of stroke. This leads to the principal question the researchers hoped to answer:

> Does the use of stents reduce the risk of stroke?

The researchers who asked this question conducted an experiment with 451 at-risk patients.[1] Each volunteer patient was randomly assigned to one of two groups:

- **Treatment group.** Patients in the treatment group received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle modification.
- **Control group.** Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

Researchers randomly assigned 224 patients to the treatment group and 227 to the control group. In this study, the control group provides a reference point against which we can measure the medical impact of stents in the treatment group.

[1] Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. *New England Journal of Medicine* 365:993–1003.

1.1.1 Examining the data

Researchers studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment. Patient outcomes were recorded as "stroke" or "no event," representing whether the patient had a stroke during that time period.

It would be difficult to answer the research question by looking at individual patient records one at a time. Instead, we organize the results into a summary table that lets us quickly compare the two groups.

Table 1.1: Descriptive statistics for the stent study.

	30 days: Stroke	30 days: No event	365 days: Stroke	365 days: No event
Treatment	33	191	45	179
Control	13	214	28	199
Total	46	405	73	378

Of the 224 patients in the treatment group, 45 had a stroke by the end of the first year. Using these two numbers, compute the proportion of patients in the treatment group who had a stroke by the end of their first year. (Note: answers to all Guided Practice exercises are provided in footnotes!)

Show answer

The proportion of the 224 patients who had a stroke within 365 days: $45/224 = 0.20 = 20\%$.

1.1.2 Drawing conclusions from the data

We can compute **sample statistics** from Table 1.1 to give us a better idea of how the impact of the stent treatment differed between the two groups.

A **sample statistic** is a single number that summarizes data from a sample. Sample statistics condense a large amount of information into a form that is easy to digest and compare.

For instance, the primary results of the study after one year could be described by two sample statistics: the proportion of people who had a stroke in the treatment and control groups.

- Proportion who had a stroke in the treatment (stent) group: $45/224 = 0.20 = 20\%$.
- Proportion who had a stroke in the control group: $28/227 = 0.12 = 12\%$.

These two sample statistics are useful in looking for differences between the groups, and we are in for a surprise: an additional 8% of patients in the treatment group had a stroke! This is important for two reasons. First, it is contrary to what doctors expected, which was that stents would *reduce* the rate of strokes. Second, it leads to a statistical question: do the data show a "real" difference between the groups?

This second question is subtle. Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably will not observe exactly 50 heads. This type of variation is part of almost any type of data generating process. It is possible that the 8% difference in the stent study is due to this natural variation. However, the larger the difference we observe (for a particular sample size), the less believable it is that the difference is due to chance. So what we are really asking is the following: if in fact stents have no effect, how unlikely is it that we would observe such a large difference?

While we do not yet have the statistical tools to fully address this question on our own, we can state the conclusion of the published analysis: there was compelling evidence of harm by stents in this study of stroke patients.

Be careful about overgeneralizing. Do not generalize the results of this study to all patients and all stents. This study looked at patients with very specific characteristics who volunteered to be part of the study and who may not be representative of all stroke patients. In addition, there are many types of stents, and this study considered only the self-expanding Wingspan stent (Boston Scientific). However, the study does leave us with an important lesson: we should keep our eyes open for surprises.

Does this study prove that stents *cause* strokes? What features of the study design might strengthen or weaken this conclusion?

Show answer

Because patients were *randomly assigned* to treatment and control groups, this is a randomized experiment. Random assignment helps ensure the groups are similar in all respects except the treatment itself. This strengthens the case for a causal conclusion. However, we should still ask: could the difference (20% vs. 12%) have arisen by chance alone? We will develop tools to answer this question rigorously later in the course.

This case study highlights a key principle that will guide much of our work in this course: **a single study's results can be surprising, and statistics gives us the tools to evaluate whether an observed result reflects a genuine effect or could be due to random chance**. We will return to this idea again and again as we develop the tools of statistical inference.

1.2 Data basics

Effective presentation and description of data is a first step in most analyses. This section introduces one structure for organizing data as well as some terminology that will be used throughout the course.

1.2.1 Observations, variables, and data matrices

An **observation** (also called a **case** or **unit of observation**) is a single entity in a dataset — one person, one county, one transaction. A **variable** is a characteristic that is measured or recorded for each observation. A **data frame** (or **data matrix**) organizes observations as rows and variables as columns. Each cell holds a single value.

Consider data for 50 randomly sampled loans offered through a peer-to-peer lending company. Each row in the table represents a single loan — that is the observation, or unit of observation. The columns represent characteristics of each loan, and each column is a variable. For example, the first row might represent a loan of $22,000 with an interest rate of 10.90%, where the borrower is based in New Jersey (NJ) and has an income of $59,000.

Table 1.2: Six observations from a loan dataset.

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	22,000	10.90	60	B	NJ	59,000	rent
2	6,000	9.92	36	B	CA	60,000	rent
3	25,000	26.30	36	E	SC	75,000	mortgage
4	6,000	9.92	36	B	CA	75,000	rent
5	25,000	9.43	60	B	OH	254,000	mortgage
6	6,400	9.92	36	B	IN	67,000	mortgage

In practice, it is especially important to ask clarifying questions to ensure important aspects of the data are understood. For instance, it is always important to be sure we know what each variable means and its units of measurement. Table 1.3 describes each variable in the loan dataset.

Table 1.3: Variables and their descriptions for the loan dataset.

Variable	Description
loan_amount	Amount of the loan received, in US dollars.
interest_rate	Interest rate on the loan, in an annual percentage.
term	The length of the loan, which is always set as a whole number of months.
grade	Loan grade, which takes values A through G and represents the quality of the loan and its likelihood of being repaid.
state	US state where the borrower resides.
total_income	Borrower's total income, including any second income, in US dollars.
homeownership	Indicates whether the person owns, owns but has a mortgage, or rents.

What is the grade of the first loan in Table 1.2? And what is the homeownership status of the borrower for that first loan?

Show answer

The loan's grade is B, and the borrower rents their residence.

1.2.2 Tidy data

A data frame where each row is a unique case (observational unit), each column is a variable, and each cell is a single value is commonly referred to as **tidy data**.

Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is tidy when:

1. Each row represents a single observation.
2. Each column represents a single variable.
3. Each cell contains a single value.

When recording data, use a tidy data frame unless you have a very good reason to use a different structure. This structure allows new cases to be added as rows or new variables as new columns and facilitates visualization, summarization, and other statistical analyses.

The grades for assignments, quizzes, and exams in a course are often recorded in a gradebook that takes the form of a data frame. How might you organize a course's grade data using a data frame? Describe the observational units and variables.

Show answer

There are multiple strategies that can be followed. One common strategy is to have each student represented by a row, and then add a column for each assignment, quiz, or exam. Under this setup, it is easy to review a single line to understand the grade history of a student. There should also be columns to include student information, such as one column to list student names.

1.2.3 A larger dataset: US counties

We consider data for 3,142 counties in the United States, which includes the name of each county, the state where it resides, its population in 2017, the population change from 2010 to 2017, poverty rate, and several additional characteristics.

How might these data be organized in a data frame? What are the observations and variables?

Show answer

Each county may be viewed as a case, and there are multiple pieces of information recorded for each case. A table with 3,142 rows and 14 columns could hold these data, where each row represents a county and each column represents a particular piece of information.

Table 1.4: Six observations and six variables from a US county dataset.

name	state	pop2017	pop_change	unemployment_rate	median_edu
Autauga County	Alabama	55,504	1.48	3.86	some_college
Baldwin County	Alabama	212,628	9.19	3.83	some_college
Barbour County	Alabama	25,270	-6.22	6.44	hs_diploma
Bibb County	Alabama	22,668	0.73	4.39	hs_diploma
Blount County	Alabama	58,013	0.68	4.02	hs_diploma
Bullock County	Alabama	10,309	-2.28	7.72	hs_diploma

Table 1.5: Variables and their descriptions for the county dataset.

Variable	Description
name	Name of county.
state	Name of state.
pop2000	Population in 2000.
pop2010	Population in 2010.
pop2017	Population in 2017.
pop_change	Population change from 2010 to 2017 (in percent).
poverty	Percent of population in poverty in 2017.
homeownership	Homeownership rate, 2006–2010.
multi_unit	Percent of housing units that are in multi-unit structures, 2006–2010.
unemployment_rate	Unemployment rate in 2017.
metro	Whether the county contains a metropolitan area (yes or no).
median_edu	Median education level (2013–2017): below_hs, hs_diploma, some_college, or bachelors.
per_capita_income	Per capita (per person) income (2013–2017).
median_hh_income	Median household income.
smoking_ban	Type of county-level smoking ban in place in 2010: none, partial, or comprehensive.

1.3 Types of variables

Examine the variables `unemployment_rate`, `pop2017`, `state`, and `median_edu` in the county dataset. Each of these variables is inherently different from the other three, yet some share certain characteristics. Understanding these distinctions is fundamental to choosing the right analysis tools.

1.3.1 Numerical variables

A **quantitative variable** (also called a **numerical variable**) takes on values where arithmetic operations like adding, subtracting, or averaging make sense.

Consider `unemployment_rate`, which can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values. On the other hand, we would not classify a variable reporting telephone area codes as quantitative since the average, sum, and difference of area codes does not have any clear meaning.

Quantitative variables can be further subdivided:

- **Continuous variables** can take any value within a range, including decimals. The `unemployment_rate` is continuous — it could be 3.86%, 4.127%, or any value in between.
- **Discrete variables** can only take counting values (typically whole numbers: 0, 1, 2, ...). The variable `pop2017` (population count) is discrete — you cannot have 55,504.7 people.

The key test for quantitative vs. categorical. Ask yourself: does it make sense to compute the average of this variable? If you can meaningfully average the values, it is quantitative. If averaging does not make sense (as with zip codes or phone numbers), the variable is categorical, even if it happens to be recorded using digits.

1.3.2 Categorical variables

A **categorical variable** (also called a **qualitative variable**) takes on values that are category names or labels. The possible values of a categorical variable are called its **levels**.

The variable `state` can take up to 51 values (after accounting for Washington, DC): Alabama, Alaska, ..., and Wyoming. Because the responses themselves are categories, `state` is a categorical variable, and the possible values (states) are the variable's levels.

Categorical variables can be further subdivided:

- **Nominal variables** have categories with no natural ordering. The variable `state` is nominal — there is no inherent reason to rank Alabama above Alaska or Wyoming.
- **Ordinal variables** have categories with a meaningful order. The variable `median_edu` takes values `below_hs`, `hs_diploma`, `some_college`, or `bachelors`. These categories have a natural ordering (less education to more education), making `median_edu` ordinal.

In most of this course, we will treat categorical variables as nominal (unordered). The distinction between nominal and ordinal becomes more important in advanced courses when special methods are used to take advantage of the ordering.

1.3.3 Variable type hierarchy

The breakdown of variable types is summarized in the diagram below.

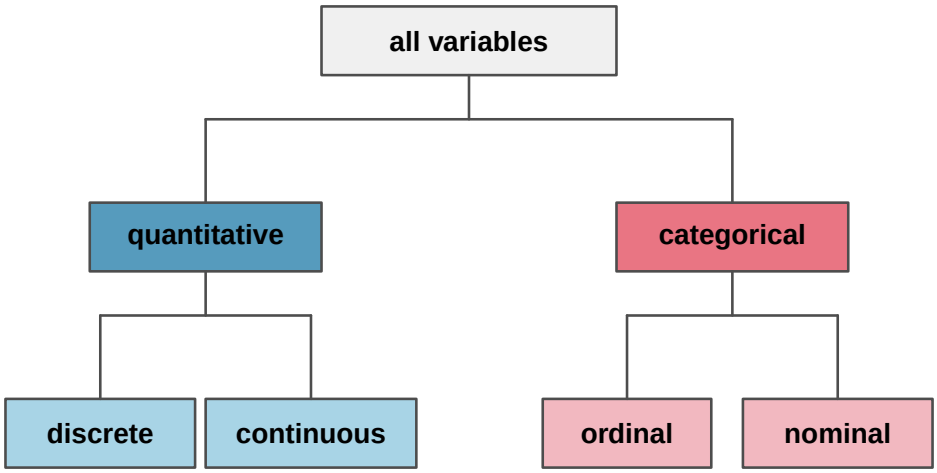


Figure 1.1: Breakdown of variables into their respective types. All variables are either quantitative or categorical. Quantitative variables are further classified as discrete or continuous. Categorical variables are further classified as ordinal or nominal.

Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous quantitative, discrete quantitative, or categorical.

The number of siblings and student height represent quantitative variables. Because the number of siblings is a count, it is discrete. Height varies continuously, so it is a continuous quantitative variable. The last variable classifies students into two categories — those who have and those who have not taken a statistics course — which makes this variable categorical.

An experiment is evaluating the effectiveness of a new drug in treating migraines. A `group` variable is used to indicate the experiment group for each patient: treatment or control. The `num_migraines` variable represents the number of migraines the patient experienced during a 3-month period. Classify each variable as either quantitative or categorical.

Show answer

The `group` variable can take just one of two group names, making it categorical. The `num_migraines` variable describes a count of the number of migraines, which is an outcome where basic arithmetic is sensible, which means this is a quantitative outcome; more specifically, since it represents a count, `num_migraines` is a discrete quantitative variable.

Classify each variable as continuous quantitative, discrete quantitative, nominal categorical, or ordinal categorical:

a. Number of text messages sent per day
b. Favorite type of music
c. Temperature in Fahrenheit
d. Satisfaction rating (1 = very unsatisfied, ..., 5 = very satisfied)

Show answer

a. Discrete quantitative — it is a count. b. Nominal categorical — the categories (rock, jazz, classical, etc.) have no natural ordering. c. Continuous quantitative — temperature can take any value in a range. d. Ordinal categorical — the numbers have a meaningful order, but the intervals between them may not be equal, and averaging satisfaction ratings is debatable.

1.4 Relationships between variables

Many analyses are motivated by a researcher looking for a relationship between two or more variables. A social scientist might like to answer questions such as:

> Does a higher-than-average increase in county population tend to correspond to counties with higher or lower median household incomes?
>
> If homeownership in one county is lower than the national average, will the percent of housing units that are in multi-unit structures in that county tend to be above or below the national average?
>
> How much can the median education level explain the median household income for counties in the US?

To answer these questions, data must be collected, such as the county dataset shown in Table 1.4. Examining sample statistics can provide numerical insights about the specifics of each of these questions. Alternatively, graphs can be used to visually explore the data, potentially providing more insight than a sample statistic.

1.4.1 Scatterplots

A **scatterplot** is a graph that displays the relationship between two quantitative variables. Each observation is represented by a point, with one variable on the horizontal axis and the other on the vertical axis.

The figure below displays the relationship between `homeownership` and `multi_unit` (the percent of housing units that are in multi-unit structures such as apartments and condos) for US counties. Each point on the plot represents a single county. For instance, one highlighted point corresponds to Chattahoochee County, Georgia, which has 39.4% of housing units in multi-unit structures and a homeownership rate of 31.3%.

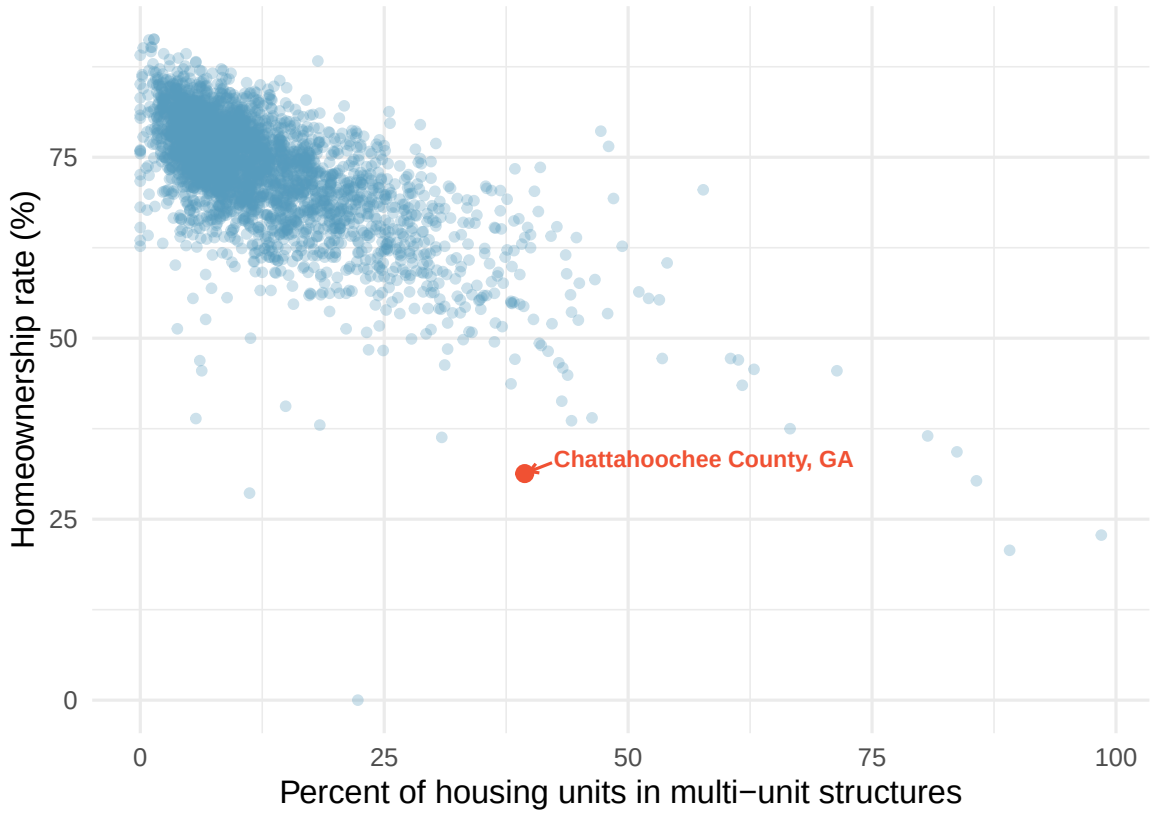


Figure 1.2: A scatterplot of homeownership rate versus the percent of housing units in multi-unit structures for US counties. There is a clear negative trend: as the multi-unit rate increases, homeownership tends to decrease. Chattahoochee County, Georgia is highlighted. Open in StatLens

The scatterplot suggests a relationship between the two variables: counties with a higher rate of housing units in multi-unit structures tend to have lower homeownership rates. We might brainstorm as to why this relationship exists and investigate each idea to determine which are the most reasonable explanations.

1.4.2 Associated and independent variables

When two variables show some connection or pattern with one another, they are called **associated** variables (or **related** variables). If two variables are not associated — there is no evident relationship between them — they are said to be **independent**.

Because there is a downward trend in the homeownership vs. multi-unit scatterplot — counties with more multi-unit housing are associated with lower homeownership — these variables are said to be **negatively associated**.

A **positive association** means that as one variable increases, the other tends to increase as well. A **negative association** means that as one variable increases, the other tends to decrease.

Consider the relationship between the percent change in population from 2010 to 2017 and median household income for counties, as shown in the scatterplot below. Are these variables associated?

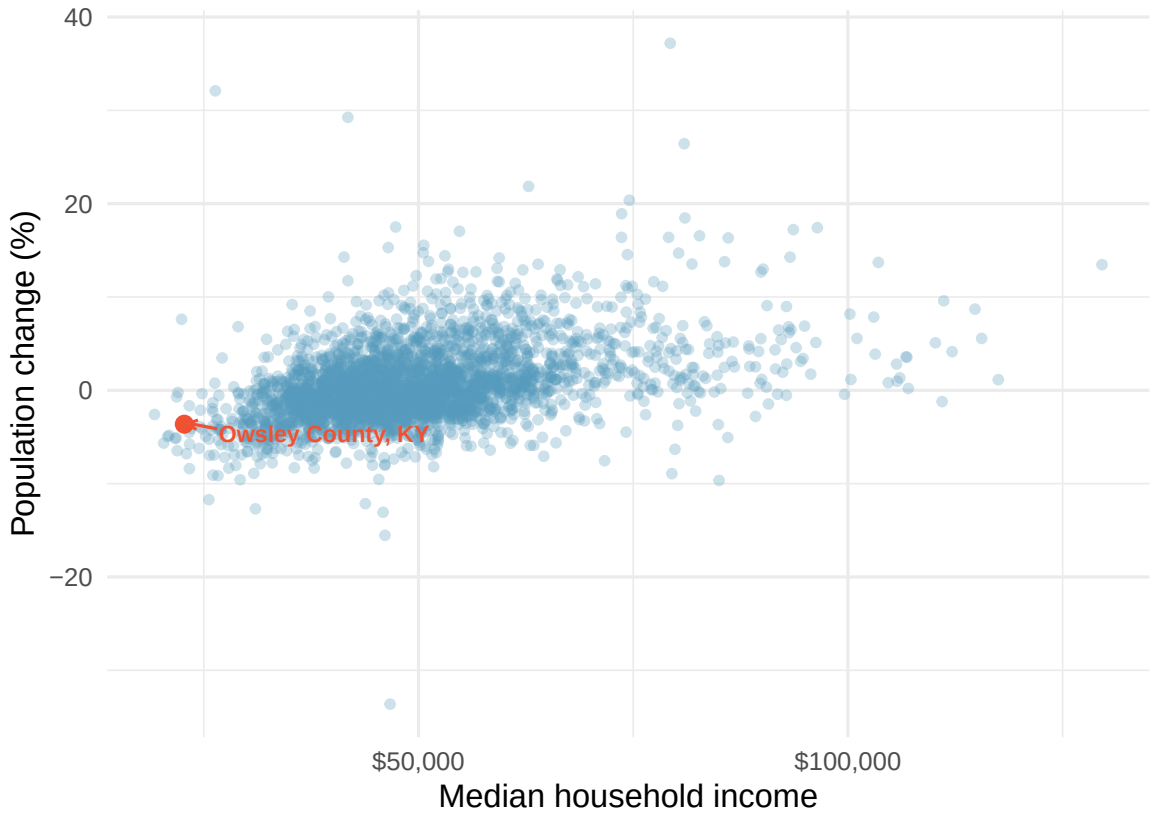


Figure 1.3: A scatterplot showing population change against median household income for US counties. There is a positive trend: counties with higher median household income tend to have higher population growth. Owsley County, Kentucky is highlighted. Open in StatLens

---

The larger the median household income for a county, the higher the population growth observed for the county. While it is not true that every county with a higher median household income has higher population growth, the trend in the plot is evident. Since there is some relationship between the variables, they are associated. Because both variables tend to increase together, this is a positive association.

Associated or independent, not both. A pair of variables are either related in some way (associated) or not (independent). No pair of variables is both associated and independent.

Examine the variables in the loan dataset, which are described in Table 1.3. Create two questions about possible relationships between variables in the loan data that are of interest to you.

Show answer

Two example questions: (1) What is the relationship between loan amount and total income? (2) If someone's income is above the average, will their interest rate tend to be above or below the average?

1.5 Explanatory and response variables

When we ask questions about the relationship between two variables, we sometimes also want to determine if the change in one variable causes a change in the other. Consider the following rephrasing of an earlier question about the county dataset:

> If there is an increase in the median household income in a county, does this *drive* an increase in its population?

In this question, we are asking whether one variable affects another.

When we suspect one variable might causally affect another, we label the first variable the **explanatory variable** and the second the **response variable**.

$$\text{explanatory variable} \longrightarrow \text{might affect} \longrightarrow \text{response variable}$$

We also use the terms explanatory and response to describe variables where the response might be *predicted* using the explanatory variable, even if there is no causal relationship.

For many pairs of variables, there is no hypothesized relationship, and these labels would not be applied to either variable.

In the stent study from Section 1.1, identify the explanatory and response variables.

The explanatory variable is the treatment group assignment (stent or control), because this is the factor that the researchers manipulated. The response variable is whether the patient had a stroke, because this is the outcome the researchers measured to evaluate the treatment's effect.

Labeling one variable as "explanatory" does not guarantee that it *causes* changes in the response. A formal evaluation of whether one variable causes changes in another requires an experiment. We discuss this distinction further in Section 1.6.

Some disciplines use the terms **independent variable** and **dependent variable** instead of explanatory and response. We avoid this language because the word "independent" has a different technical meaning in statistics — it refers to variables that are not associated. Using "independent" in two different ways can cause confusion.

A researcher is studying whether hours spent studying for an exam are related to exam scores. Identify the explanatory and response variables.

Show answer

The explanatory variable is hours spent studying, because this is the factor that might influence the outcome. The response variable is exam score, because this is the outcome being measured.

1.6 Observational studies and experiments

There are two primary types of data collection: experiments and observational studies. Understanding the difference is critical, because the type of study determines what conclusions we can draw.

1.6.1 Experiments

Studies where the researchers assign treatments to cases are called **experiments**. When individuals are randomly assigned to treatment groups, the experiment is called a **randomized experiment**.

For instance, we may suspect that drinking a high-calorie energy drink will improve performance in a race. To check if there really is a causal relationship between the explanatory variable (whether the runner drank an energy drink or not) and the response variable (the race time), researchers identify a sample of individuals and split them into groups. The individuals in each group are *assigned* a treatment. Random assignment organizes the participants into groups that are roughly equal on all

aspects, thus allowing us to control for any other variables that might affect the outcome (such as fitness level, racing experience, or age).

For example, each runner in the experiment could be randomly assigned — perhaps by flipping a coin — into one of two groups: the first group receives a **placebo** (a fake treatment, in this case a no-calorie drink that looks identical to the energy drink) and the second group receives the high-calorie energy drink.

A **placebo** is a fake treatment that is made to look like the real treatment. Placebos are used so that subjects in the control group have the same experience as those in the treatment group (except for the actual treatment), which helps researchers isolate the effect of the treatment itself.

See the stent case study in Section 1.1 for another example of an experiment, though that study did not employ a placebo.

1.6.2 Observational studies

Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise. In an observational study, the researcher merely observes what happens naturally.

Researchers perform observational studies when they collect data via surveys, review medical or company records, or follow a **cohort** (a group of similar individuals) over time to form hypotheses about why certain outcomes might occur. In each of these situations, researchers merely observe the data that arise — they do not assign treatments to subjects.

A researcher wants to know if smoking causes lung cancer. She follows 10,000 adults over 20 years, recording whether each person smokes and whether they develop lung cancer. Is this an experiment or an observational study?

This is an observational study. The researcher did not assign anyone to smoke or not smoke — she simply observed the participants' behavior and health outcomes. While the study might reveal an association between smoking and lung cancer, it cannot by itself prove causation, because there may be other differences between smokers and non-smokers (such as diet, exercise habits, or genetics) that contribute to the outcome.

1.6.3 The key distinction: association vs. causation

In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection. This is because in observational studies, researchers cannot control for all the other variables that might be influencing the outcome.

Association does not imply causation.

In general, association does not imply causation. An advantage of a randomized experiment is that it is easier to establish causal relationships, because random assignment tends to balance out the effects of other variables across treatment groups. Establishing causation from observational studies requires advanced statistical methods that are beyond the scope of this course. We revisit these ideas when we discuss study design in the study design chapter.

Determine whether each of the following is an experiment or an observational study:

a. A pharmaceutical company randomly assigns 200 patients to receive either a new drug or a placebo and measures their cholesterol levels after 6 months.
b. A university surveys its graduates to determine whether those who participated in study-abroad programs earn higher salaries five years after graduation.

c. A teacher assigns students in her morning class to use a new learning app and students in her afternoon class to study using traditional methods, then compares exam scores.

---

a. This is an experiment — the company randomly assigns patients to treatments.
b. This is an observational study — the university simply collects data on students who chose (or chose not) to study abroad. The students were not randomly assigned.
c. This is tricky. The teacher assigns groups, so it might seem like an experiment. However, because students were not randomly assigned (they ended up in morning vs. afternoon classes for their own reasons), this is better described as a quasi-experiment or an observational study. Differences in outcomes could be due to the app, or they could be due to systematic differences between the types of students who take morning vs. afternoon classes.

1.6.4 Putting it all together

The table below summarizes the key differences between experiments and observational studies.

	Experiment	Observational Study
Researcher's role	Assigns treatments	Observes what happens naturally
Random assignment?	Yes (in a randomized experiment)	No
Can establish causation?	Yes, if randomized (for subjects like those in the study)	Generally no
Example	Randomly assign patients to drug vs. placebo	Survey people about their exercise habits and health

1.7 Chapter review

1.7.1 Summary

This chapter introduced you to the world of data. Here are the key ideas:

- **Data** are observations collected about the world around us. **Statistics** is the study of how best to collect, analyze, and draw conclusions from data.
- A **data frame** organizes data so that each row represents an observation and each column represents a variable. Data organized this way is called **tidy data**.
- Variables come in two fundamental types: **quantitative** (continuous or discrete) and **categorical** (nominal or ordinal).
- When two variables show a pattern together, they are **associated**; if not, they are **independent**. Associations can be **positive** (both increase together) or **negative** (one increases while the other decreases).
- The **explanatory variable** is the variable we think might influence the **response variable**.
- **Randomized experiments** randomly assign subjects to treatments and can establish causation for subjects like those in the study. **Observational studies** merely observe what happens and can identify associations, but generally cannot establish causation.
- **Association does not imply causation** — this is one of the most important principles in statistics.

Many of the ideas from this chapter will be revisited as we move on to doing end-to-end data analyses. In the next chapter, you will learn about how we can design studies to collect the data we need to make conclusions with the desired scope of inference.

1.7.2 Key terms

The following terms were introduced in this chapter. You should be able to define and give an example of each.

Term	Term	Term
associated	data	data frame
case	categorical variable	cohort
continuous	dependent variable	discrete
experiment	explanatory variable	independent
level	negative association	nominal
quantitative variable	observation	observational study
ordinal	placebo	positive association
randomized experiment	response variable	sample statistic
tidy data	unit of observation	variable

1.8 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Marvel Cinematic Universe films.** The data frame below contains information on Marvel Cinematic Universe films through the Infinity saga (a movie storyline spanning from Ironman in 2008 to Endgame in 2019). Box office totals are given in millions of US Dollars. How many observations and how many variables does this data frame have?

		Length				Gross	
	Title	Hrs	Mins	Release Date	Opening Wknd US	US	World
1	Iron Man	2	6	5/2/2008	98.62	319.03	585.8
2	The Incredible Hulk	1	52	6/12/2008	55.41	134.81	264.77
3	Iron Man 2	2	4	5/7/2010	128.12	312.43	623.93
4	Thor	1	55	5/6/2011	65.72	181.03	449.33
5	Captain America: The First Avenger	2	4	7/22/2011	65.06	176.65	370.57
...	...	...	...	...	...	...	...
23	Spiderman: Far from Home	2	9	7/2/2019	92.58	390.53	1131.93

2. **Cherry Blossom Run.** The data frame below contains information on runners in the 2017 Cherry Blossom Run, which is an annual road race that takes place in Washington, DC. Most runners participate in a 10-mile run while a smaller fraction take part in a 5k run or walk. How many observations and how many variables does this data frame have?

						Time			
	Bib	Name	Sex	Age	City / Country	Net	Clock	Pace	Event
1	6	Hiwot G.	F	21	Ethiopia	3217	3217	321	10 Mile
2	22	Buze D.	F	22	Ethiopia	3232	3232	323	10 Mile
3	16	Gladys K.	F	31	Kenya	3276	3276	327	10 Mile
4	4	Mamitu D.	F	33	Ethiopia	3285	3285	328	10 Mile
5	20	Karolina N.	F	35	Poland	3288	3288	328	10 Mile
...	...	...	...	...	...	...	...	...	...
19961	25153	Andres E.	M	33	Woodbridge, VA	5287	5334	1700	5K

3. **Air pollution and birth outcomes, study components.** Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study, air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and inhalable particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births. (Ritz et al. 2000)

a. Identify the main research question of the study.

b. Who are the subjects in this study, and how many are included?

c. What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

4. **Cheaters, study components.** Researchers studying the relationship between honesty, age and self-control conducted an experiment on 160 children between the ages of 5 and 15. Participants reported their age, sex, and whether they were an only child or not. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. (Bucciol and Piovesan 2011)

a. Identify the main research question of the study.

b. Who are the subjects in this study, and how many are included?

c. The study's findings can be summarized as follows: *"Half the students were explicitly told not to cheat and the others were not given any explicit instructions. In the no instruction group probability of cheating was found to be uniform across groups based on child's characteristics. In the group that was explicitly told to not cheat, girls were less likely to cheat, and while rate of cheating didn't vary by age for boys, it decreased with age for girls."* How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

5. **Gamification and statistics, study components.** Gamification is the application of game-design elements and game principles in non-game contexts. In educational settings, gamification is often implemented as educational activities to solve problems by using characteristics of game elements. Researchers investigating the effects of gamification on learning statistics conducted a study where they split college students in a statistics class into four groups: (1) no reading exercises and no gamification, (2) reading exercises but no gamification, (3) gamification but no reading exercises, and (4) gamification and reading exercises. Students in all groups also attended lectures. Students in the class were from two majors: Electrical and Computer Engineering (n = 279) and Business Administration (n = 86). After their assigned learning experience, each student took a final evaluation comprised of 30 multiple choice question and their score was measured as the number of questions they answered correctly. The researchers considered students' gender, level of studies (first through fourth year) and academic major. Other variables considered were expertise in the English language and use of personal computers and games, both of which were measured on a scale of 1 (beginner) to 5 (proficient). The study found that gamification had a positive effect on student learning compared to traditional teaching methods involving lectures and reading exercises. They also found that the effect was larger for females and Engineering students. (Legaki et al. 2020)

a. Identify the main research question of the study.

b. Who were the subjects in this study, and how many were included?

c. What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

6. **Stealers, study components.** In a study of the relationship between socio-economic class and unethical behavior, 129 University of California undergraduates at Berkeley were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken. (Piff et al. 2012)

a. Identify the main research question of the study.

b. Who were the subjects in this study, and how many were included?

c. The study found that students who were identified as upper-class took more candy than others. How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

7. **Migraine and acupuncture.** A migraine is a particularly painful type of headache, which patients sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a randomized controlled study where 89 individuals who identified as female diagnosed with migraine headaches were randomly assigned to one of two groups: treatment or control. Forty-three (43) patients in the treatment group received acupuncture that is specifically designed to treat migraines. Forty-six (46) patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). Twenty-four (24) hours after patients received acupuncture, they were asked if they were pain free. Results are summarized in the contingency table below. Also provided is a figure from the original paper displaying the appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks. (Allais et al. 2011)

	Pain free?	
Group	No	Yes
Control	44	2
Treatment	33	10

a. What percent of patients in the treatment group were pain free 24 hours after receiving acupuncture?

b. What percent were pain free in the control group?

c. In which group did a higher percent of patients become pain free 24 hours after receiving acupuncture?

d. Your findings so far might suggest that acupuncture is an effective treatment for migraines for all people who suffer from migraines. However this is not the only possible conclusion. What is one other possible explanation for the observed difference between the percentages of patients that are pain free 24 hours after receiving acupuncture in the two groups?

e. What are the explanatory and response variables in this study?

8. **Sinusitis and antibiotics.** Researchers studying the effect of antibiotic treatment for acute sinusitis compared to symptomatic treatments randomly assigned 166 adults diagnosed with acute sinusitis to one of two groups: treatment or control. Study participants received either a 10-day course of amoxicillin (an antibiotic) or a placebo similar in appearance and taste. The placebo consisted of symptomatic treatments such as acetaminophen, nasal decongestants, etc. At the end of the 10-day period, patients were asked if they experienced improvement in symptoms. The distribution of responses is summarized below. (Garbutt et al. 2012)

	Improvement	
Group	No	Yes
Control	16	65
Treatment	19	66

a. What percent of patients in the treatment group experienced improvement in symptoms?

b. What percent experienced improvement in symptoms in the control group?

c. In which group did a higher percentage of patients experience improvement in symptoms?

d. Your findings so far might suggest a real difference in the effectiveness of antibiotic and placebo treatments for improving symptoms of sinusitis. However this is not the only possible conclusion. What is one other possible explanation for the observed difference between the percentages patients who experienced improvement in symptoms?

e. What are the explanatory and response variables in this study?

9. **Daycare fines, study components.** Researchers tested the deterrence hypothesis which predicts that the introduction of a penalty will reduce the occurrence of the behavior subject to the fine, with the condition that the fine leaves everything else unchanged, by instituting a fine for late pickup at daycare centers. For this study, they worked with 10 volunteer daycare centers that did not originally impose a fine to parents for picking up their kids late. They randomly selected 6 of these daycare centers and instituted a monetary fine (of a considerable amount) for picking up children late and then removed it. In the remaining 4 daycare centers no fine was introduced. The study period was divided into four: before the fine (weeks 1–4), the first 4 weeks with the fine (weeks 5-8), the last 8 weeks with fine (weeks 9–16), and the after fine period (weeks 17-20). Throughout the study, the number of kids who were picked up late was recorded each week for each daycare. The study found that the number of late-coming parents increased discernibly when the fine was introduced, and no reduction occurred after the fine was removed. (Gneezy and Rustichini 2000)

center	week	group	late_pickups	study_period
1	1	test	8	before fine
1	2	test	8	before fine
1	3	test	7	before fine
1	4	test	6	before fine
1	5	test	8	first 4 weeks with fine
...	...	...	...	...
10	20	control	13	after fine

   a. Is this an observational study or an experiment? Explain your reasoning.

   b. What are the cases in this study and how many are included?

   c. What is the response variable in the study and what type of variable is it?

   d. What are the explanatory variables in the study and what types of variables are they?

10. **Efficacy of COVID-19 vaccine on adolescents, study components.** Results of a Phase 3 trial announced in March 2021 show that the Pfizer-BioNTech COVID-19 vaccine demonstrated 100% efficacy and robust antibody responses on 12 to 15 years old adolescents with or without prior evidence of SARS-CoV-2 infection. In this trial 2,260 adolescents were randomly assigned to two groups: one group got the vaccine (n = 1,131) and the other got a placebo (n = 1,129). While 18 cases of COVID-19 were observed in the placebo group, none were observed in the vaccine group. (Pfizer 2021)

   a. Is this an observational study or an experiment? Explain your reasoning.

   b. What are the cases in this study and how many are included?

   c. What is the response variable in the study and what type of variable is it?

   d. What are the explanatory variables in the study and what types of variables are they?

11. **Palmer penguins.** Data were collected on 344 penguins living on three islands (Torgersen, Biscoe, and Dream) in the Palmer Archipelago, Antarctica. In addition to which island each penguin lives on, the data contains information on the species of the penguin (*Adelie*, *Chinstrap*, or *Gentoo*), its bill length, bill depth, and flipper length (measured in millimeters), its body mass (measured in grams), and the sex of the penguin (female or male). (Gorman, Williams, and Fraser 2014)

   a. How many cases were included in the data?

b. How many numerical variables are included in the data? Indicate what they are, and if they are continuous or discrete.
c. How many categorical variables are included in the data, and what are they? List the corresponding levels (categories) for each.

12. **Smoking habits of UK residents.** A survey was conducted to study the smoking habits of 1,691 UK residents. Below is a data frame displaying a portion of the data collected in this survey. A blank cell indicates that data for that variable was not available for a given respondent.

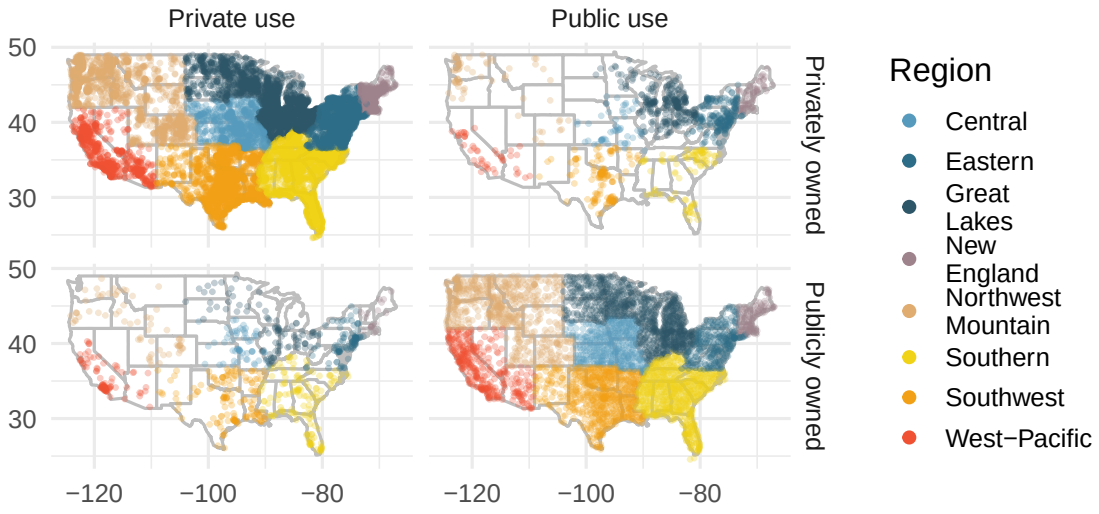
						amount	
	sex	age	marital_status	gross_income	smoke	weekend	weekday
1	Female	61	Married	2,600 to 5,200	No	NA	NA
2	Female	61	Divorced	10,400 to 15,600	Yes	5	4
3	Female	69	Widowed	5,200 to 10,400	No	NA	NA
4	Female	50	Married	5,200 to 10,400	No	NA	NA
5	Male	31	Single	10,400 to 15,600	Yes	10	20
...	...	...	...	...	...	NA	NA
1691	Male	49	Divorced	Above 36,400	Yes	15	10

a. What does each row of the data frame represent?

b. How many participants were included in the survey?

c. Indicate whether each variable is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

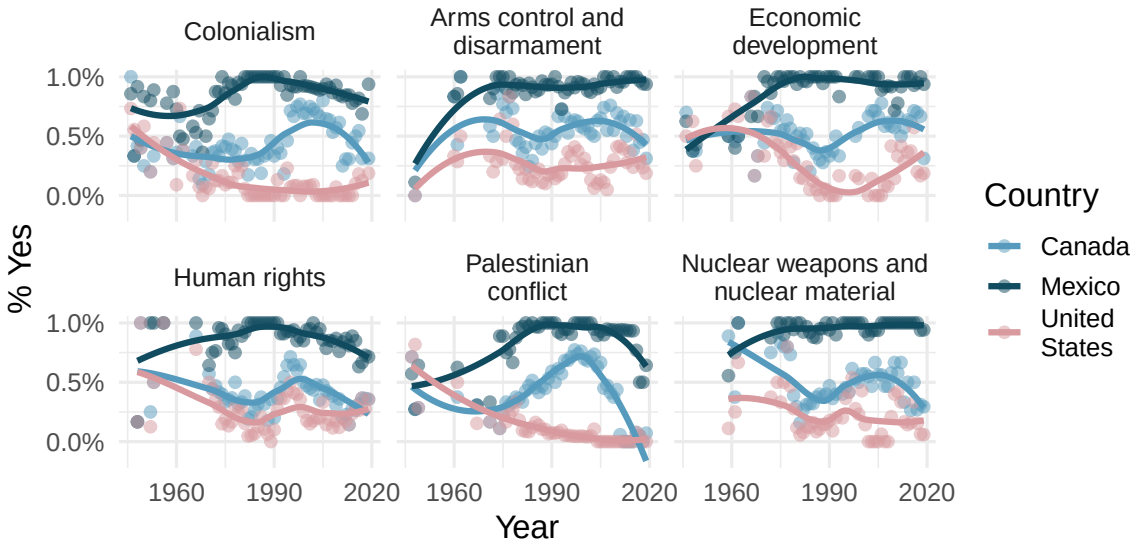
13. **US Airports.** The visualization below shows the geographical distribution of airports in the contiguous United States and Washington, DC. This visualization was constructed based on a dataset where each observation is an airport.



a. List the variables you believe were necessary to create this visualization.

b. Indicate whether each variable is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

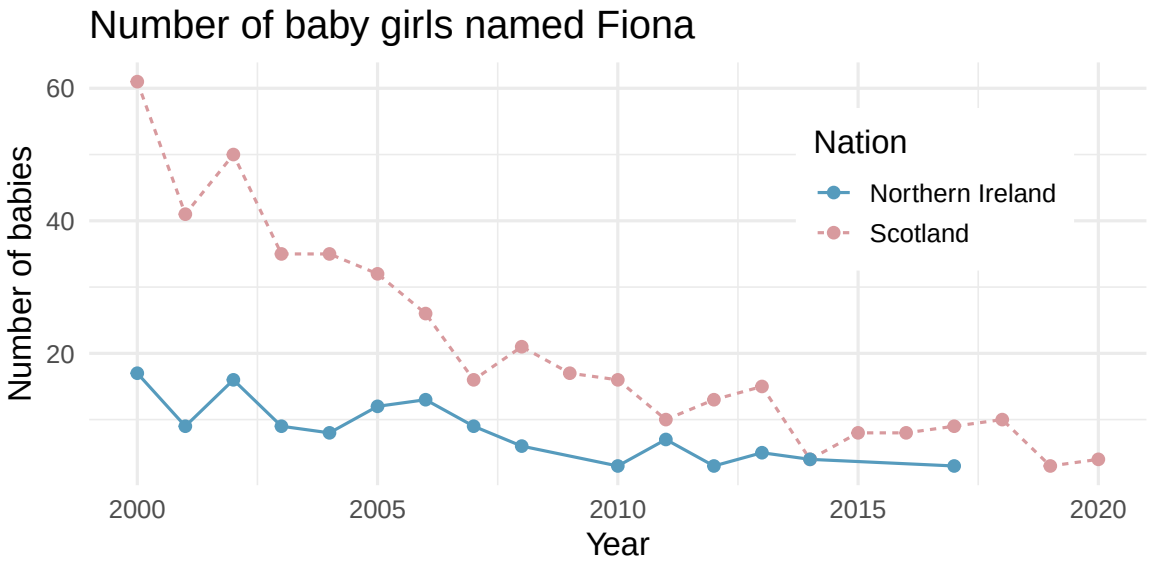
14. **UN Votes.** The visualization below shows voting patterns in the United States, Canada, and Mexico in the United Nations General Assembly on a variety of issues. Specifically, for a given year between 1946 and 2019, it displays the percentage of roll calls in which the country voted yes for each issue. This visualization was constructed based on a dataset where each observation is a country/year pair.



a. List the variables used in creating this visualization.

b. Indicate whether each variable is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

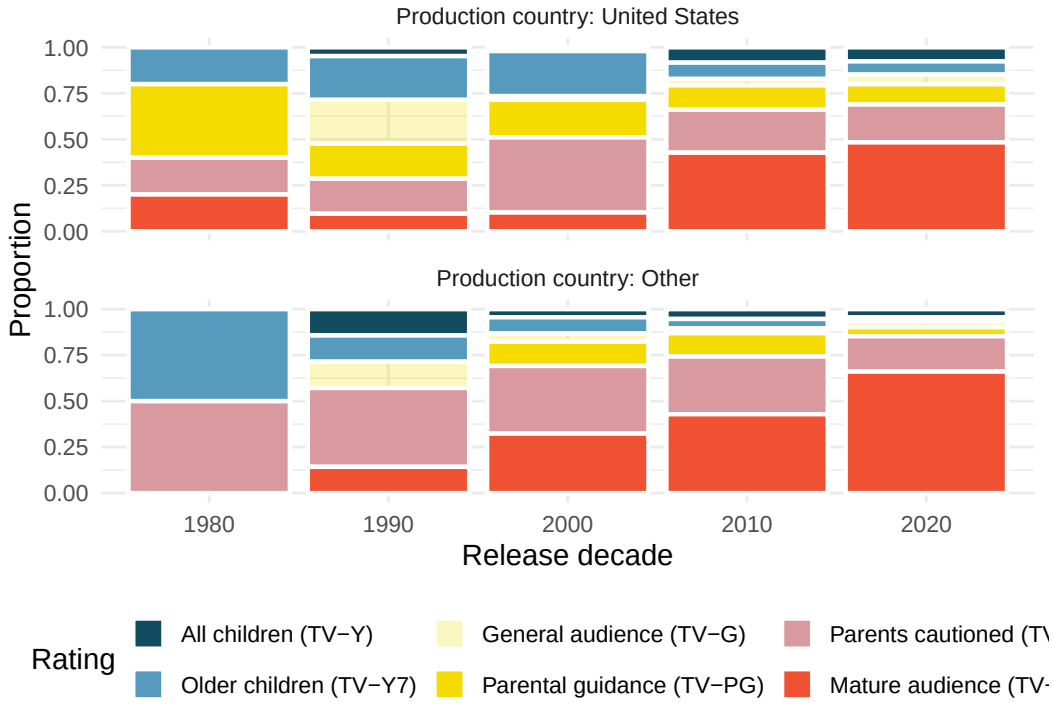
15. **UK baby names.** The visualization below shows the number of baby girls born in the United Kingdom (comprised of England & Wales, Northern Ireland, and Scotland) who were given the name "Fiona" over the years.



a. List the variables you believe were necessary to create this visualization.

b. Indicate whether each variable is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

16. **Shows on Netflix.** The visualization below shows the distribution of ratings of TV shows on Netflix (a streaming entertainment service) based on the decade they were released in and the country they were produced in. In the dataset, each observation is a TV show.



a. List the variables you believe were necessary to create this visualization.

b. Indicate whether each variable is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

17. **Stanford Open Policing.** The Stanford Open Policing project gathers, analyzes, and releases records from traffic stops by law enforcement agencies across the United States. Their goal is to help researchers, journalists, and policy makers investigate and improve interactions between police and the public. The following is an excerpt from a summary table created based off of the data collected as part of this project. (Pierson et al. 2020)

		Driver		Car	
County	State	Race / Ethnicity	Arrest rate	Stops / year	Search rate
Apache County	AZ	Black	0.016	266	0.077
Apache County	AZ	Hispanic	0.018	1008	0.053
Apache County	AZ	White	0.006	6322	0.017
Cochise County	AZ	Black	0.015	1169	0.047
...	...	...	...	...	...
Wood County	WI	Hispanic	0.029	27	0.036
Wood County	WI	White	0.029	1157	0.033

a. What variables were collected on each individual traffic stop in order to create the summary table above?

b. State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.

c. Suppose we wanted to evaluate whether vehicle search rates are different for drivers of different races. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

18. **Space launches.** The following summary table shows the number of space launches in the US by the type of launching agency and the outcome of the launch (success or failure).

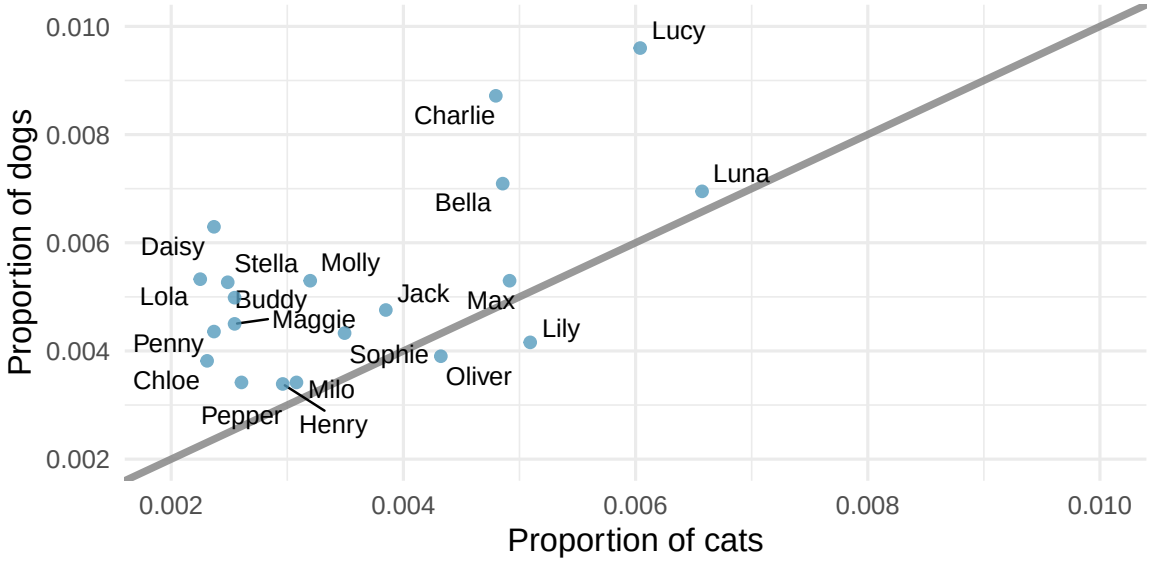
	1957 - 1999		2000-2018	
	Failure	Success	Failure	Success
Private	13	295	10	562
State	281	3751	33	711
Startup	0	0	5	65

a. What variables were collected on each launch in order to create to the summary table above?

b. State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.

c. Suppose we wanted to study how the success rate of launches vary between launching agencies and over time. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

19. **Pet names.** The city of Seattle, WA has an open data portal that includes pets registered in the city. For each registered pet, we have information on the pet's name and species. The following visualization plots the proportion of dogs with a given name versus the proportion of cats with the same name. The 20 most common cat and dog names are displayed. The diagonal line on the plot is the $x = y$ line; if a name appeared on this line, the name's popularity would be exactly the same for dogs and cats.



a. Are these data collected as part of an experiment or an observational study?

b. What is the most common dog name? What is the most common cat name?

c. What names are more common for cats than dogs?

d. Is the relationship between the two variables positive or negative? What does this mean in context of the data?

20. **Stressed out in an elevator.** In a study evaluating the relationship between stress and muscle cramps, half the subjects are randomly assigned to be exposed to increased stress by being placed into an elevator that falls rapidly and stops abruptly and the other half are left at no or baseline stress.

a. What type of study is this?

b. Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

Datasets: `mcu_films` (openintro) | `run17` (cherryblossom) | `migraine` (openintro) | `sinusitis` (openintro) | `daycare_fines` (openintro) | `biontech_adolescents` (openintro) | `smoking` (openintro) | `usairports` (airports) | `ukbabynames` (ukbabynames) | `netflix_titles` (tidytuesdayR) | `seattlepets` (openintro)

Chapter 2

Data Collection and Study Design

Before digging into the details of working with data, we pause to think about how data come to be. If data are to be used to draw broad conclusions, then it is important to understand who or what the data represent. One important aspect is **sampling** — knowing how the observational units were selected from a larger group allows us to generalize back to the population from which the data were drawn. Additionally, by understanding the structure of the study, we can separate causal relationships from mere associations. A good question to ask before analyzing any dataset is: *How were these observations collected?* You will learn a lot about the data by understanding its source.

StatLens: Sampling Bias Lab. Convinced you could pick a "representative" sample by eye? Try the Sampling Bias Lab. You hand-pick "representative" words from the Gettysburg Address (the population), then the tool draws random samples of the same size. Both sets of sample means are plotted against the true population mean ($\mu \approx 4.3$ letters). The point of the demo: human-chosen samples are systematically biased, and the bias does **not** shrink as n grows. Other sampling methods exist, and plenty of real studies use them — but random sampling is the one that lets sample statistics estimate population parameters honestly.

2.1 Populations and samples

The first step in conducting research is to identify the question to be investigated. A clearly stated research question helps identify what subjects or cases should be studied and what variables are important. It is also important to consider *how* data are collected so that they are reliable and help achieve the research goals.

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last five years, what is the average time to complete a degree for undergraduates at a public university?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

The **population** is the entire group of individuals or cases about which we want information. A **sample** is a subset of the population from which we actually collect data. A **census** is a study that attempts to collect data from every member of the population.

Each research question refers to a target population. In the first question, the target population is all swordfish in the Atlantic Ocean, and each fish represents a case. Oftentimes it is not feasible to collect data for every case in a population. A census is difficult because it may be too expensive, or it might be difficult or impossible to even identify every member of the population. Instead, we take a

sample. Ideally, a sample is a manageable fraction of the population. For instance, 60 swordfish might be selected, and this sample data may be used to estimate the population average mercury content.

For the second and third research questions above, identify the target population and what represents an individual case.

Show answer

For the question about time to complete a degree, the population includes all undergraduates at the university who graduated in the last five years (we can only compute an average for students who actually finished). Each graduate is an individual case. For the question about the heart disease drug, the population includes all people with severe heart disease, and each person represents a case.

2.1.1 Parameters and statistics

In most statistical analyses, the research question boils down to determining a numerical summary — perhaps a quantity you already know (like the average) or one you will learn about in this course.

A numerical summary can be calculated on either the sample or the entire population. However, measuring every unit in the population is usually impossible. So we calculate the summary from a sample and use it to estimate the corresponding population value.

A **population parameter** is a numerical summary calculated from (or defined for) the entire population. A **sample statistic** is a numerical summary calculated from the sample data.

We use specific terms to distinguish between numbers calculated from a sample (sample statistics) and numbers that describe the entire population (population parameters). These terms will be used extensively in later chapters when we discuss making inferences about populations.

A poll of 1,000 likely voters finds that 54% plan to vote for Candidate A. Identify the population parameter and the sample statistic.

Show answer

The sample statistic is 54% — this is the proportion observed in the sample of 1,000 voters. The population parameter is the true proportion of *all* likely voters who plan to vote for Candidate A, which is unknown. The sample statistic (54%) is our best estimate of that unknown parameter.

2.1.2 Anecdotal evidence

Consider the following possible responses to the three research questions from above:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 7 years to graduate from the university, so it must take longer to graduate here than at many other schools.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Each conclusion is based on data. However, there are two problems. First, the data only represent one or two cases. Second, and more importantly, it is unclear whether these cases are actually representative of the population. Data collected in this haphazard fashion are called **anecdotal evidence**.

Be careful of anecdotal evidence. Such evidence may be true and verifiable, but it may only represent extraordinary cases and therefore not be a good representation of the population.

Anecdotal evidence typically consists of unusual cases that we recall based on their striking characteristics. For instance, we are more likely to remember the two people who took 7 years to graduate

than the six others who graduated in four years. Instead of focusing on the most unusual cases, we should examine a sample of many cases that better represent the population.

2.2 Sampling from a population

We might try to estimate the time to graduation for the university's undergraduates by collecting a sample of graduates. All graduates in the last five years represent the *population*, and graduates who are selected for review are collectively called the *sample*. In general, we always seek to *randomly* select a sample from a population. The most basic type of random selection is equivalent to how raffles are conducted. For example, we could write each graduate's name on a raffle ticket and draw 10 tickets. The selected names would represent a random sample of 10 graduates.

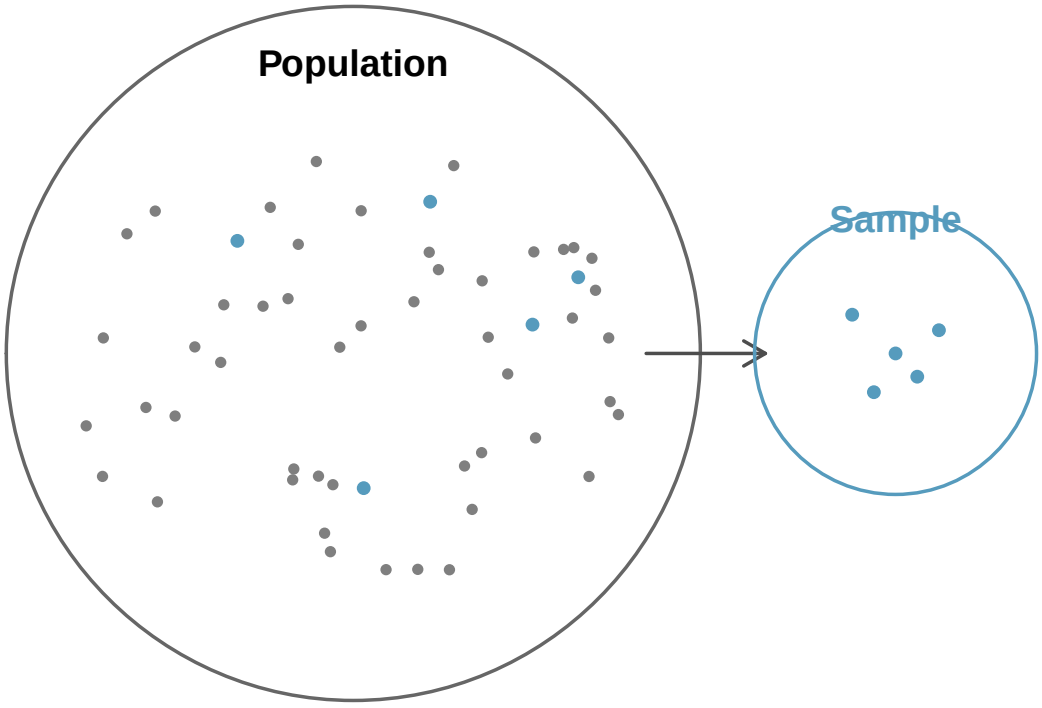


Figure 2.1: Sampling from a population. Individuals are randomly selected from the population (large circle) to be included in the sample (small circle).

Suppose we ask a student who happens to be majoring in nutrition to select several graduates for the study. Which students do you think they might pick? Do you think their sample would be representative of all graduates?

They might pick a disproportionate number of graduates from health-related fields. When selecting samples by hand, we run the risk of picking a **biased** sample, even if our bias is unintentional.



Figure 2.2: A biased sample. A nutrition major asked to select graduates might inadvertently pick a disproportionate number from health-related fields.

If someone is permitted to pick and choose exactly which individuals are included in the sample, it is entirely possible that the sample will overrepresent that person's interests — even if the bias is unintentional. Sampling randomly helps address this problem.

A **simple random sample** (**SRS**) is a sample drawn so that each case in the population has an equal chance of being included, and the selection of any one case is independent of the selection of any other case. This is equivalent to drawing names out of a hat.

The act of taking a simple random sample helps minimize bias. However, bias can still creep in, and it helps to separate two quite different ways that happens. Some biases come from *how the sample was chosen*; others arise *after* the sample is chosen, from how people respond or how the response is measured.

Sampling bias arises from the way cases are selected: the selection mechanism systematically over- or under-represents part of the population. **Non-sampling bias** arises after selection — from who actually responds, from how questions are asked, or from how a measurement is taken. A perfectly drawn random sample can still be undone by non-sampling bias.

2.2.1 Sampling bias

One common form is a **convenience sample**, where individuals who are easily accessible are more likely to be included. For instance, if a political survey is conducted by stopping people on a single street corner, the results will not represent the broader city. It is often difficult to discern what sub-population a convenience sample actually represents.

A second form is **voluntary response bias**, where the sample consists of people who selected *themselves* into it. Online polls, call-in surveys, and product reviews all work this way, and the people motivated to take part are rarely a cross-section of everyone who could have.

We can easily access ratings for products, sellers, and companies on websites like Amazon. These ratings come only from people who go out of their way to provide a rating. If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product? Why or why not?

Show answer

Probably not. People tend to be more motivated to leave a review when they have a negative experience than when a product meets expectations. This creates a negative bias in product ratings — dissatisfied customers are overrepresented among reviewers. This is **voluntary response bias** — a sampling bias, because the people who end up in the sample selected themselves, and they differ systematically from the buyers who stayed silent.

2.2.2 Non-sampling bias

Even a genuine random sample can mislead. Suppose people are selected at random for a survey — for example, by dialing random phone numbers — but many of them never answer. If only 30% of the people sampled actually respond, then it is unclear whether the results are **representative** of the entire population. This **non-response bias** can skew results, because the 30% who reply may differ systematically from the 70% who do not.

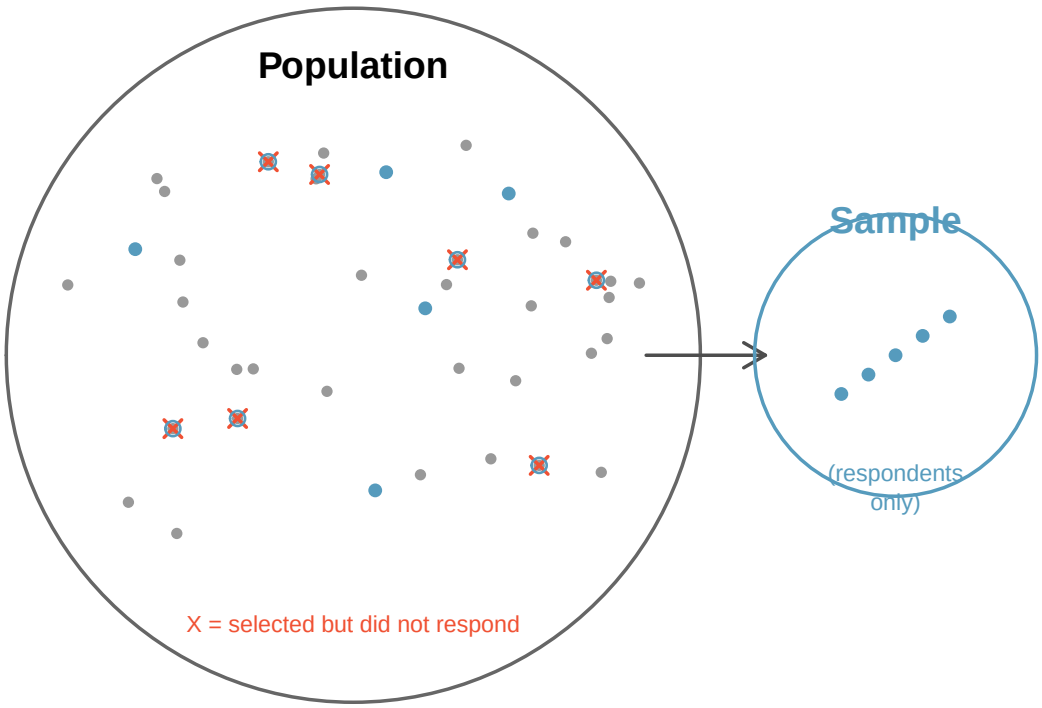


Figure 2.3: Non-response bias. Survey studies may only reach a certain group within the population. Non-response can make even a random sample unrepresentative.

Non-response is not the only way this happens. **Response bias** occurs when the measurement itself distorts the answer — a leading question ("Don't you agree that ...?"), a sensitive topic people under-report, or an interviewer whose presence changes what respondents are willing to say. Neither problem is fixed by sampling more carefully; both are fixed by asking better and following up harder.

2.3 Random sampling methods

Almost all statistical methods are based on the notion of randomness. If data are not collected using a random process from a population, then the estimates and the errors associated with those estimates are not reliable. Here we consider four random sampling techniques: simple random sampling, stratified sampling, cluster sampling, and multistage sampling.

2.3.1 Simple random sampling

Simple random sampling is the most intuitive form of random sampling. Consider the salaries of Major League Baseball (MLB) players, where each player belongs to one of the league's 30 teams. To take a simple random sample of 120 baseball players and their salaries, we could write the names of all players onto slips of paper, drop the slips into a bucket, mix them thoroughly, then draw out 120 slips. In general, a sample is called "simple random" if each case in the population has an equal chance of being included in the final sample *and* knowing that a particular case is included tells us nothing about which other cases are included. (This population is actually small — there are fewer than 1,000 MLB players — so a full **census** would be feasible here; we use it only because the setting is familiar.)

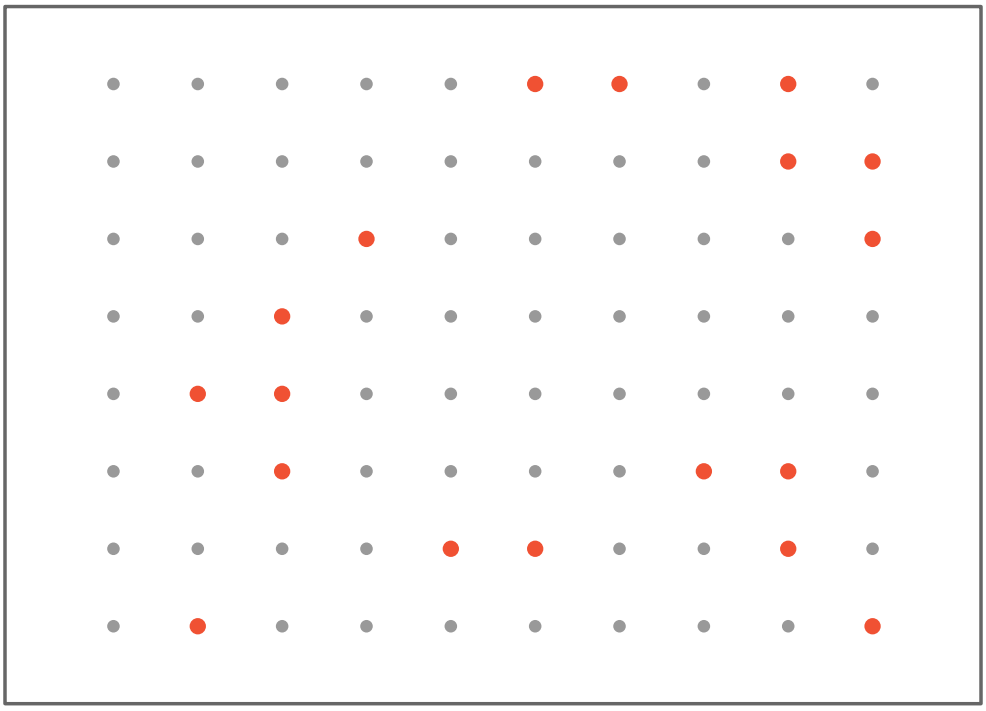


Figure 2.4: Simple random sampling. 18 cases (shown in red) are randomly selected from the population.

2.3.2 Stratified sampling

Stratified sampling is a divide-and-conquer strategy. The population is divided into groups called **strata**. The strata are chosen so that similar cases are grouped together, then a second sampling method (usually simple random sampling) is used within each stratum. In the baseball salary example, each of the 30 teams could represent a stratum, since some teams have much larger payrolls than others. We might randomly sample 4 players from each team to get our sample of 120.

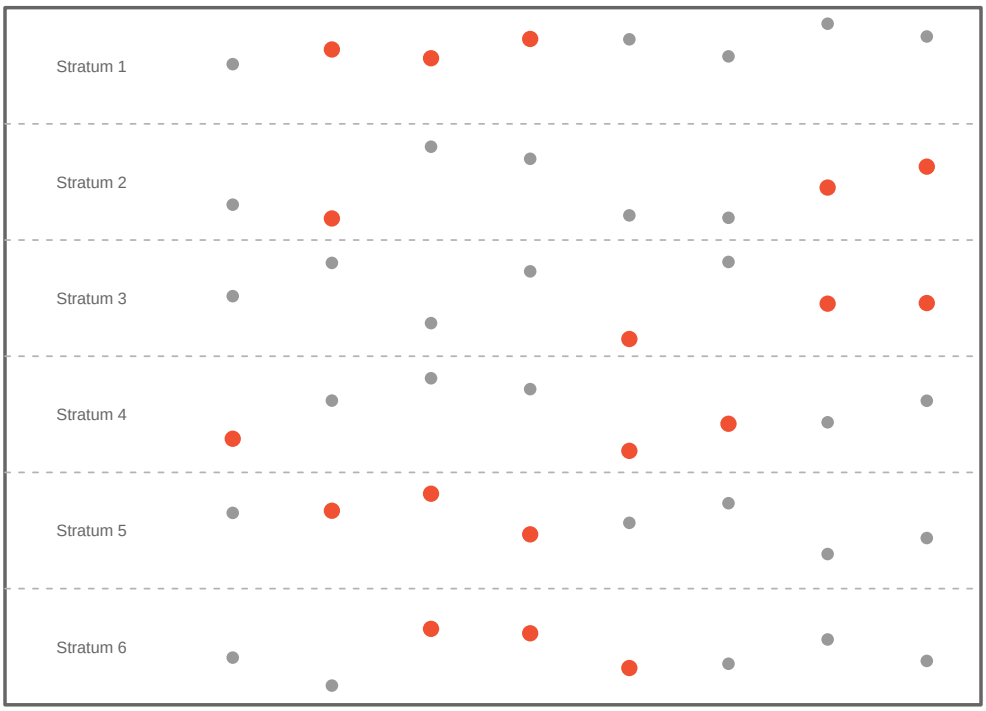


Figure 2.5: Stratified sampling. Cases are first grouped into strata, then simple random sampling is used within each stratum to select 3 cases per stratum.

Stratified sampling is especially useful when cases within each stratum are very similar with respect to the outcome of interest.

Why is it beneficial for cases within each stratum to be very similar?

If cases within a stratum are similar, we can get a very stable (precise) estimate for each subgroup. When we combine these stratum-level estimates into a single estimate for the full population, the overall estimate will tend to be more precise because each individual component is itself more precise.

2.3.3 Cluster sampling

In a **cluster sample**, we break up the population into many groups called **clusters**, then randomly select a fixed number of clusters and include *all* observations from each selected cluster in the sample.

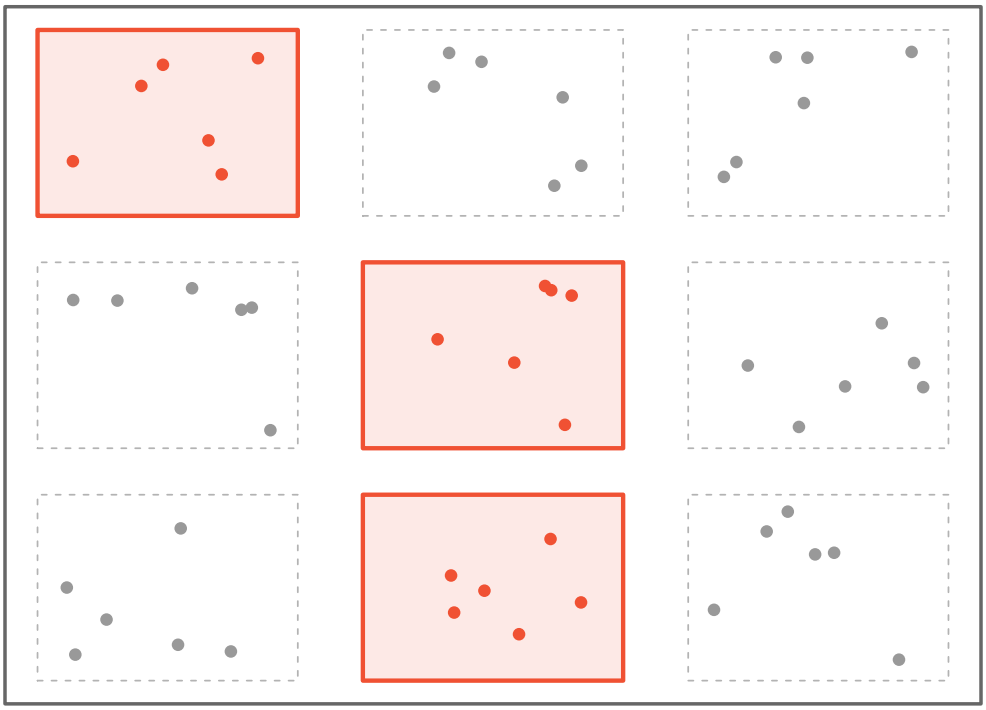


Figure 2.6: Cluster sampling. The population is divided into nine clusters. Three clusters are randomly selected, and all observations within those clusters are included in the sample.

2.3.4 Multistage sampling

A **multistage sample** is like a cluster sample, but instead of keeping all observations in each selected cluster, we take a random sample *within* each selected cluster.

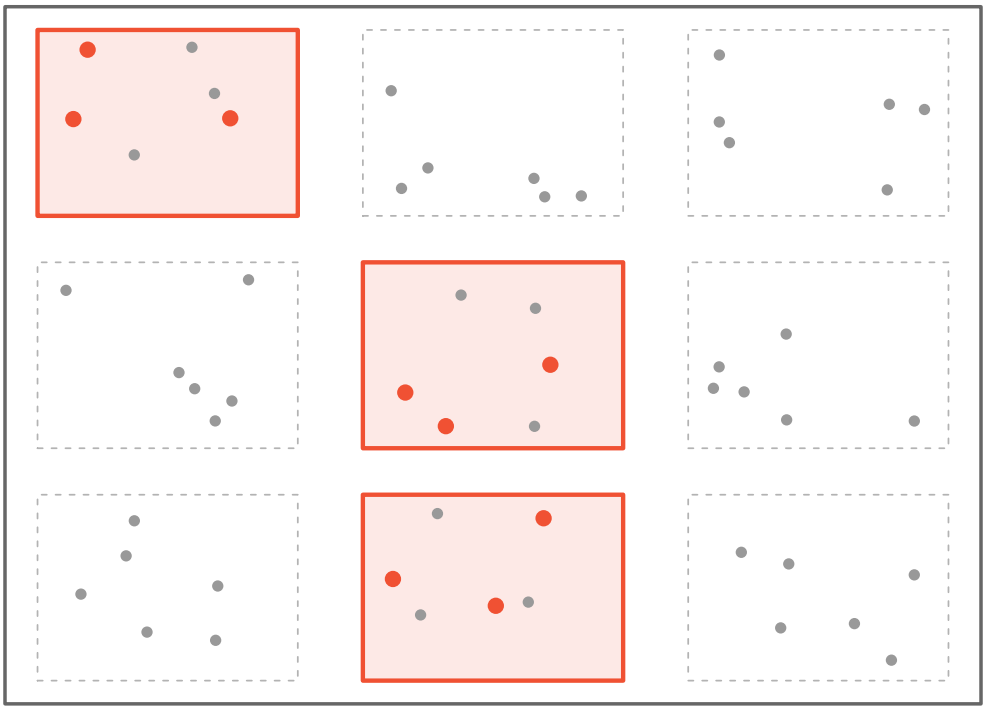


Figure 2.7: Multistage sampling. Three clusters are randomly selected, then a random sample of observations is drawn from within each selected cluster.

Cluster and multistage sampling can be more economical than other techniques. Unlike stratified sampling, these approaches work best when there is a lot of case-to-case variability *within* a cluster but the clusters themselves are fairly similar to one another. For example, if neighborhoods represent clusters, then cluster sampling works best when the residents within each neighborhood are diverse.

Suppose we want to estimate the malaria rate in a densely tropical portion of rural Indonesia. There are 30 villages in the region, each more or less like the next, but the distances between villages are substantial. We want to test 150 individuals for malaria. What sampling method should we use?

A simple random sample would likely draw individuals from all 30 villages, making data collection very expensive given the travel distances. Stratified sampling would be difficult because it is unclear how to group individuals into similar strata. However, cluster or multistage sampling are excellent choices. With multistage sampling, we could randomly select, say, 15 of the 30 villages, then randomly select 10 people from each village. This would substantially reduce data collection costs while still yielding reliable information.

2.3.5 Comparing the four methods

The table below summarizes the key features of each sampling method:

Method	How it works	Best used when...
Simple random	Every case has an equal chance of selection	Population is relatively homogeneous and accessible
Stratified	Divide into similar groups (strata), then SRS within each	Known subgroups differ on the outcome of interest
Cluster	Randomly select entire groups (clusters)	Clusters are internally diverse but similar to each other; travel/cost is a concern
Multistage	Randomly select clusters, then SRS within each	Same as cluster, but clusters are large

2.4 Experiments

An **experiment** is a study where the researchers actively assign treatments to cases. A **treatment** is a specific condition the researcher imposes on cases — one value of the explanatory variable (for example, a particular drug or dose). When this assignment uses randomization (e.g., a coin flip to decide which treatment a patient receives), it is called a **randomized experiment**. Randomized experiments are fundamentally important when trying to establish a causal connection between two variables.

2.4.1 Principles of experimental design

Well-designed experiments incorporate four key principles:

1. **Controlling.** Researchers assign treatments to cases and do their best to **control** any other differences between the groups. For example, when patients take a drug in pill form, some patients might take the pill with a sip of water while others drink a full glass. To control for the effect of water consumption, a doctor might instruct every patient to drink a 12-ounce glass of water with the pill.

2. **Randomization.** Researchers randomly assign subjects to treatment groups to account for variables that cannot be controlled. For example, some patients may be more susceptible to a disease than others due to their dietary habits. Randomly assigning patients to treatment or control groups helps even out such differences across groups.

3. **Replication.** The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we **replicate** by collecting a sufficiently large sample. At a minimum, we want multiple subjects per treatment group. Replication also refers to repeating an entire study to verify earlier findings. (The **replication crisis** refers to the ongoing problem in several scientific disciplines where past findings have failed to be replicated in new studies.)

4. **Blocking.** Researchers sometimes know or suspect that variables other than the treatment influence the response. They may first group individuals based on this variable into **blocks** and then randomize cases within each block to the treatment groups.[1]

[1] Why *blocking*? The term comes from agricultural experiments, where researchers literally divided a field into relatively homogeneous **blocks** of plots and randomized treatments within each block. It is a useful coincidence that the procedure also "blocks out" some nuisance variation, but that does not appear to be the historical origin of the word. (*Suggested by Todd Will.*) For instance, in studying the effect of a drug on heart attacks, researchers might first split patients into low-risk and high-risk blocks, then randomly assign half the patients from each block to the treatment and half to the control group.

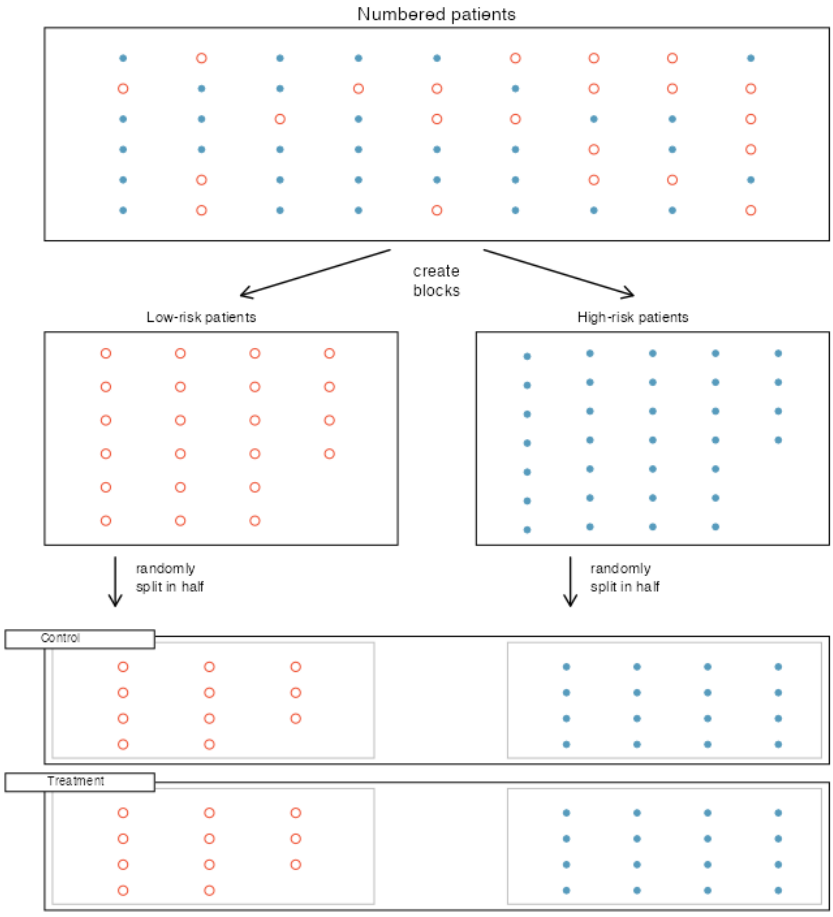


Figure 2.8: Blocking for patient risk. Patients are first divided into low-risk and high-risk blocks, then patients in each block are evenly randomized into treatment groups.

It is important to incorporate at least the first three principles into any study. Blocking is a slightly more advanced technique that adds precision when relevant grouping variables are known. Blocking is the experimental cousin of **stratified sampling**: both first group similar units together — strata when *selecting* a sample, blocks when *assigning* a treatment — so that a known source of variation is handled deliberately rather than left to chance.

Matched pairs are a common special case of blocking where each block has exactly two units. The most familiar version uses the *same subject* measured twice — a before/after design, or a crossover where each participant receives both treatments in random order. A different form pairs *distinct* subjects who match on a relevant variable (twins, littermates, students matched on baseline test score) and then randomizes the two treatments within each pair. A typical example: participants complete a mental-rotation test, get an hour of instruction, then take a parallel post-test — each person's *change*

score isolates the effect of instruction from their individual baseline ability. In both forms the pair acts as its own control, absorbing subject-to-subject variability that would otherwise make a real treatment effect harder to detect. (Less of that noise means the estimated effect varies less from one study to the next — an idea we develop in the sampling variability chapter.) We revisit matched-pairs data — and the paired analysis it enables — in the confidence intervals for means chapter and the hypothesis tests for means chapter.

2.4.2 Reducing bias in human experiments

Randomized experiments are considered the gold standard for establishing cause-and-effect, but they do not automatically eliminate all bias. Human studies are a perfect example of where bias can unintentionally arise.

Consider a study where a new drug is tested to see if it reduces deaths in heart attack patients. Researchers designed a randomized experiment: study volunteers were randomly placed into two groups. The **treatment group** received the drug. The **control group** did not receive any drug treatment.

Put yourself in the place of a person in the study. If you are in the treatment group, you receive a promising new drug and may feel hopeful. A person in the control group receives nothing and might feel anxious. These emotional effects could influence health outcomes independently of the drug itself. This means there are actually two effects at play: the drug's real effectiveness, and the emotional impact of knowing (or not knowing) that you are receiving treatment.

To prevent emotional effects from biasing the study, researchers want patients to be unaware of their group assignment — that is, the study should be *blind*. But if a patient in the control group receives nothing at all, they will know they are not getting treated. The solution is a placebo. An effective placebo is the key to making a study truly blind.

The patients are not the only ones who should be blinded. Doctors and researchers can also unintentionally bias a study. A doctor who knows a patient is receiving the real treatment might give that patient more attention or care. To guard against this, most modern studies use a **double-blind** setup.

A study is **blind** (or **single-blind**) if the patients do not know which treatment group they are in. A **placebo** is a fake treatment (such as a sugar pill) given to the control group so that patients cannot tell whether they are receiving the real treatment. The **placebo effect** is the phenomenon where patients show improvement simply because they believe they are receiving treatment. A study is **double-blind** if neither the patients nor the doctors and researchers interacting with them know who is receiving which treatment.

Think back to the stent study described in Section 1.1. Is it an experiment? Was the study blinded? Was it double-blinded?

Show answer

The researchers assigned patients to treatment groups, so it is an experiment. However, patients could distinguish what treatment they received because a stent is a surgical procedure. There is no simple equivalent of a "placebo stent" (though sham surgeries do exist), so the study was likely not fully blinded. If the study was not blind, it could not be double-blind either.

For the stent study, could the researchers have employed a placebo? If so, what would it have looked like?

Show answer

In principle, researchers could use a **sham surgery** — the patient undergoes a surgical procedure but does not actually receive the stent. Sham surgeries do exist in medical research, but they raise

serious ethical questions because they expose control-group patients to the risks of surgery without the potential benefit of the treatment.

There are always multiple viewpoints on experiments and placebos, and rarely is it obvious which approach is ethically "correct." Is it ethical to use a sham surgery that creates risk for the patient? On the other hand, if we do not use one, we might promote a costly treatment that has no real effect, diverting resources from treatments that actually work.

2.5 Observational studies

An **observational study** is a study where no treatment has been explicitly applied or withheld by the researchers. The researchers simply observe and record what happens naturally.

Making causal conclusions based on experiments is often reasonable because researchers can randomly assign treatments. However, making causal conclusions from observational data is much more dangerous and is generally not recommended. Observational studies are typically only sufficient to show **associations** or to form hypotheses that can later be tested with experiments.

Suppose an observational study tracked sunscreen use and skin cancer, and found that people who used more sunscreen were *more* likely to have skin cancer. Does this mean sunscreen causes skin cancer?

No! Previous research tells us that sunscreen actually *reduces* skin cancer risk. The missing piece is **sun exposure**. People who spend a lot of time in the sun are more likely to use sunscreen *and* more likely to get skin cancer. Sun exposure is a **confounding variable** — it is associated with both sunscreen use and skin cancer, creating a misleading association between the two.

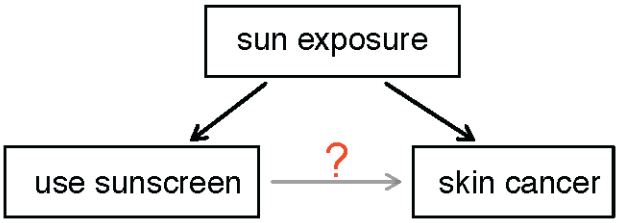


Figure 2.9: Sun exposure as a confounding variable. Sun exposure causes both increased sunscreen use and increased skin cancer risk, creating a spurious association between sunscreen use and skin cancer.

2.6 Confounding variables

A **confounding variable** is a variable that is associated with *both* the explanatory variable and the response variable. Because of this dual association, a confounding variable makes it impossible to determine whether the explanatory variable truly caused the observed response.

Confounding variables may or may not be measured as part of the study. Regardless, drawing cause-and-effect conclusions from observational studies is difficult because of the ever-present possibility of confounding.

Consider a classic example: ice cream sales and drowning rates both increase during summer months. If we looked only at the association between ice cream sales and drownings, we might (absurdly) conclude that ice cream causes drowning. The confounding variable is **temperature** (or season) —

warm weather drives both increased ice cream consumption and increased swimming, which leads to more drowning incidents.

Suppose data show a negative association between homeownership rates and the percentage of housing units that are in multi-unit structures across U.S. counties. Is it reasonable to conclude that building more apartments *causes* homeownership rates to drop? Suggest a confounding variable.

Show answer

No, a causal conclusion is not warranted from observational data. A likely confounding variable is **population density**. Dense urban counties tend to have more multi-unit housing (apartments, condos) because of limited space, and they also tend to have lower homeownership rates because property values are higher. Population density is associated with both variables.

2.6.1 Prospective and retrospective studies

Observational studies come in two forms:

- A **prospective study** identifies individuals and collects information as events unfold going forward. For example, medical researchers might identify and follow a group of patients over many years to assess the influence of behavior on cancer risk. The famous Nurses' Health Study, which started in 1976 and has collected data on over 275,000 nurses, is a prospective study.
- A **retrospective study** collects data after events have already taken place — for example, by reviewing past medical records.

Some studies contain both prospective and retrospective elements. For instance, a medical study might gather information about participants' histories (retrospective) and then follow them forward in time (prospective).

2.7 Random assignment and causation

The distinction between random *sampling* and random *assignment* is crucial. They serve different purposes and affect what conclusions we can draw.

- **Random sampling** (selecting subjects randomly from a population) allows us to **generalize** our findings to the broader population.
- **Random assignment** (randomly assigning subjects to treatment groups in an experiment) allows us to draw **causal** conclusions.

Random sampling and random assignment are independent concepts. A study can have one, both, or neither. The type of randomization determines the scope of conclusions we can draw.

A company wants to know whether a new keyboard layout increases typing speed. They recruit 40 volunteers from their office, randomly assign 20 to train on the new layout and 20 to continue with the standard layout, then measure typing speed after one month.

Can the company conclude that the new layout *causes* faster typing? Can they generalize the results to all office workers?

Because the study used **random assignment** to treatment groups, the company can make a causal conclusion: any observed difference in typing speed can be attributed to the keyboard layout rather than to pre-existing differences between the groups. However, the 40 volunteers were not randomly selected from all office workers — they are a convenience sample of volunteers from one company. Therefore, the results may not generalize to all office workers.

2.8 Scope of inference

The scope of inference — what conclusions we can draw from a study — depends on how data were collected. The following table summarizes the four possible scenarios:

	Random assignment (to treatments)	**No random assignment**
Random sampling (from population)	Can generalize to the population **and** can draw causal conclusions	Can generalize to the population, but **cannot** draw causal conclusions
No random sampling	**Cannot** generalize to the population, but can draw causal conclusions (for subjects like those in the study)	**Cannot** generalize **and cannot** draw causal conclusions

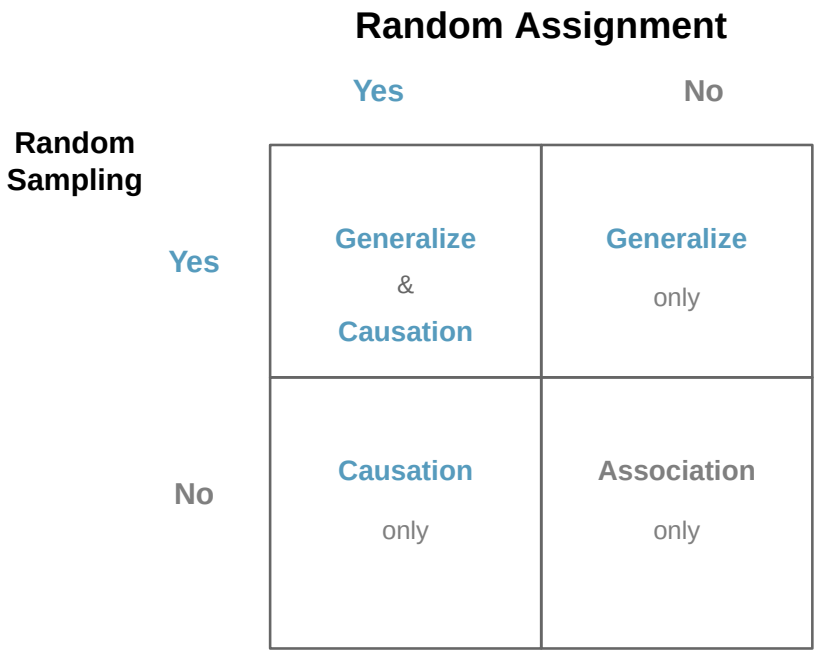


Figure 2.10: Scope of inference. The type of sampling determines whether results generalize to the population; the type of assignment determines whether causal conclusions are possible.

Very few studies achieve both random sampling *and* random assignment. In practice, experiments typically use volunteers (not random samples from the population), so they fall in the bottom-left cell: they can establish causation, but the results may not generalize beyond the type of people who volunteered. Observational studies based on random samples (top-right cell) can generalize to the population but cannot establish cause-and-effect.

A researcher randomly selects 500 adults from a city's voter rolls and surveys them about exercise habits and depression symptoms. She finds that people who exercise more report fewer depression symptoms. What conclusions can she draw?

Show answer

Because the researcher used **random sampling** (from the voter rolls), she can generalize her findings to the city's adult voter population. However, because she did not randomly **assign** people to exercise or not exercise (this is an observational study), she cannot conclude that exercise *causes* fewer depression symptoms. There may be confounding variables — for example, people with fewer depression symptoms may be more motivated to exercise in the first place.

A pharmaceutical company recruits 200 volunteers with chronic pain and randomly assigns half to receive a new pain medication and half to receive a placebo. The group receiving the medication reports substantially less pain. What is the scope of inference?

Show answer

Because the study used **random assignment**, the company can conclude that the medication *caused* the reduction in pain (a causal conclusion). However, because the participants were volunteers rather than a random sample of all chronic pain patients, the results may not generalize to the entire population of chronic pain sufferers — only to people similar to those who volunteered.

2.9 Chapter review

2.9.1 Summary

In this chapter, we explored how data are collected and why the method of collection matters for the conclusions we can draw.

Key concepts:

- The **population** is the entire group of interest; a **sample** is a subset from which we collect data. A **census** collects data on every member of the population but is usually impractical.
- **Anecdotal evidence** — conclusions drawn from one or a few non-representative cases — is unreliable.
- **Random sampling methods** include simple random sampling, stratified sampling, cluster sampling, and multistage sampling. All four use randomization — they differ in how the population is organized before the random draw. Each has advantages depending on the population's structure and practical constraints.
- **Bias** comes in two kinds. **Sampling bias** comes from *how cases are selected* — convenience samples and voluntary response. **Non-sampling bias** arises *after* selection — non-response and response bias. A perfect random sample is no protection against the second kind.
- In an **experiment**, researchers actively assign treatments to subjects. In an **observational study**, researchers simply observe without intervening.
- **Random assignment** to treatment groups is the key to establishing **causation**.
- **Confounding variables** are associated with both the explanatory and response variables and can create misleading associations.
- **Blinding** (single and double) and **placebos** help reduce bias in human experiments.
- The **scope of inference** depends on whether the study used random sampling (which enables generalization) and random assignment (which enables causal conclusions).

2.9.2 Key terms

Term	Definition
Population	The entire group of individuals about which we want information
Sample	A subset of the population from which we collect data
Census	An attempt to collect data from every member of the population
Population parameter	A numerical summary of the entire population
Sample statistic	A numerical summary calculated from sample data
Anecdotal evidence	Data collected in a haphazard fashion, often representing unusual cases
Simple random sample (SRS)	A sample where every case has an equal chance of selection, independently
Stratified sampling	Dividing the population into strata, then taking an SRS within each
Cluster sampling	Randomly selecting entire clusters and including all cases within them
Multistage sampling	Randomly selecting clusters, then taking an SRS within each selected cluster
Bias	Systematic favoritism in data collection that makes a sample unrepresentative
Sampling bias	Bias arising from *how* cases are selected into the sample
Convenience sample	A sample where easily accessible individuals are overrepresented
Voluntary response bias	Bias arising when people select themselves into the sample
Non-sampling bias	Bias arising *after* selection, from response or measurement
Non-response bias	Bias introduced when sampled individuals do not respond
Response bias	Bias introduced when the measurement itself distorts the answer
Experiment	A study where the researchers assign treatments to subjects
Randomized experiment	An experiment where treatments are assigned using randomization
Observational study	A study where researchers observe without assigning treatments
Confounding variable	A variable associated with both the explanatory and response variables
Treatment	A specific condition assigned to subjects; a value of the explanatory variable chosen by the researcher
Treatment group	The group receiving the experimental treatment
Control group	The group not receiving the treatment (or receiving a placebo)
Placebo	A fake treatment given to the control group
Placebo effect	Improvement in patients who believe they are receiving treatment
Blind (single-blind)	Subjects do not know their treatment assignment
Double-blind	Neither subjects nor researchers know the treatment assignments
Blocking	Grouping subjects by a known variable before random assignment
Prospective study	An observational study that follows subjects forward in time
Retrospective study	An observational study that looks back at past data
Scope of inference	The extent to which study results can be generalized and/or support causal claims

2.10 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Parameters and statistics.** Identify which value represents the sample mean and which value represents the claimed population mean.

a. American households spent an average of about $52 in 2007 on Halloween merchandise such as costumes, decorations and candy. To see if this number had changed, researchers conducted a new survey in 2008 before industry numbers were reported. The survey included 1,500 households and found that average Halloween spending was $58 per household.

b. The average GPA of students in 2001 at a private university was 3.37. A survey on a sample of 203 students from this university yielded an average GPA of 3.59 a decade later.

2. **Sleeping in college.** A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night. Identify which value represents the sample mean and which value represents the claimed population mean.

3. **Air pollution and birth outcomes, scope of inference.** Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. (Ritz et al. 2000)

a. Identify the population of interest and the sample in this study.

b. Comment on whether the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

4. **Cheaters, scope of inference.** Researchers studying the relationship between honesty, age and self-control conducted an experiment on 160 children between the ages of 5 and 15. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. Differences were observed in the cheating rates in the instruction and no instruction groups, as well as some differences across children's characteristics within each group. (Bucciol and Piovesan 2011)

a. Identify the population of interest and the sample in this study.

b. Comment on whether the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

5. **Gamification and statistics, scope of inference.** Researchers investigating the effects of gamification (application of game-design elements and game principles in non-game contexts) on learning statistics randomly assigned 365 college students in a statistics course to one of four groups; one of these groups had no reading exercises and no gamification, one group had reading but no gamification, one group had gamification but no reading, and a final group had gamification and reading. Students in all groups also attended lectures. The study found that gamification had a positive impact on student learning compared to traditional teaching methods involving reading exercises. (Legaki et al. 2020)

a. Identify the population of interest and the sample in this study.

b. Comment on whether the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

6. **Stealers, scope of inference.** In a study of the relationship between socio-economic class and unethical behavior, 129 University of California undergraduates at Berkeley were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken. It was found that those who were identified as upper-class took more candy than others. (Piff et al. 2012)

a. Identify the population of interest and the sample in this study.

b. Comment on whether the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

7. **Relaxing after work.** The General Social Survey asked the question, *"After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?"* to a random sample of 1,155 Americans. The average relaxing time was found to be 1.65 hours. Determine which of the following is an observation, a variable, a sample statistic, or a population parameter.

a. An American in the sample.

b. Number of hours spent relaxing after an average work day.

c. 1.65.

d. Average number of hours all Americans spend relaxing after an average work day.

8. **Cats on YouTube.** Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all videos on YouTube so you use a random video picker to select 1000 videos for you. You find that 2% of these videos are cat videos. Determine which of the following is an observation, a variable, a sample statistic, or a population parameter.

a. Percentage of all videos on YouTube that are cat videos.

b. 2%.

c. A video in your sample.

d. whether a video is a cat video.

9. **Course satisfaction across sections.** A large college class has 160 students. All 160 students attend the lectures together, but the students are divided into 4 groups, each of 40 students, for lab sections administered by different teaching assistants. The professor wants to conduct a survey about how satisfied the students are with the course, and he believes that the lab section a student is in might affect the student's overall satisfaction with the course.

a. What type of study is this?

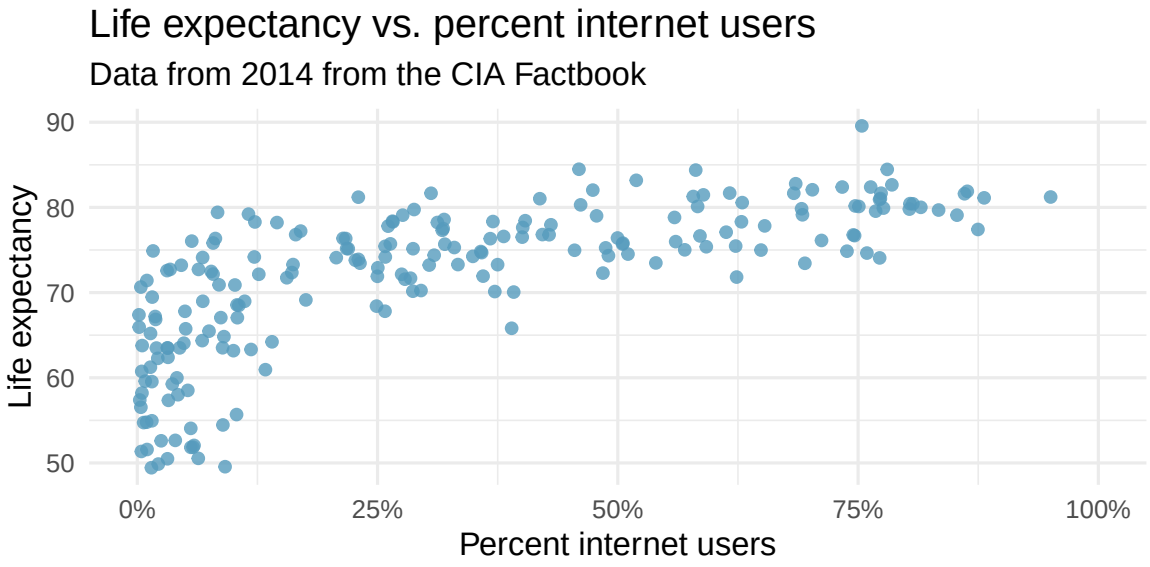
b. Suggest a sampling strategy for carrying out this study.

10. **Housing proposal across dorms.** On a large college campus first-year students and sophomores live in dorms located on the eastern part of the campus and juniors and seniors live in dorms located on the western part of the campus. Suppose you want to collect student opinions on a new housing structure the college administration is proposing and you want to make sure your survey equally represents opinions from students from all years.

a. What type of study is this?

b. Suggest a sampling strategy for carrying out this study.

11. **Internet use and life expectancy.** The following scatterplot was created as part of a study evaluating the relationship between estimated life expectancy at birth (as of 2014) and percentage of internet users (as of 2009) in 208 countries for which such data were available.



a. Describe the relationship between life expectancy and percentage of internet users.

b. What type of study is this?

c. State a possible confounding variable that might explain this relationship and describe its potential effect.

12. **Stressed out.** A study that surveyed a random sample of otherwise healthy high school students found that they are more likely to get muscle cramps when they are stressed. The study also noted that students drink more coffee and sleep less when they are stressed.

a. What type of study is this?

b. Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

c. State possible confounding variables that might explain the observed relationship between increased stress and muscle cramps.

13. **Evaluate sampling methods.** A university wants to determine what fraction of its undergraduate student body support a new $25 annual fee to improve the student union. For each proposed method below, indicate whether the method is reasonable or not.

a. Survey a simple random sample of 500 students.

b. Stratify students by their field of study, then sample 10% of students from each stratum.

c. Cluster students by their ages (e.g., 18 years old in one cluster, 19 years old in one cluster, etc.), then randomly sample three clusters and survey all students in those clusters.

14. **Random digit dialing.** The Gallup Poll uses a procedure called random digit dialing, which creates phone numbers based on a list of all area codes in America in conjunction with the associated number of residential households in each area code. Give a possible reason the Gallup Poll chooses to use random digit dialing instead of picking phone numbers from the phone book.

15. **Haters are gonna hate, study confirms.** A study published in the *Journal of Personality and Social Psychology* asked a group of 200 randomly sampled participants recruited online using Amazon's Mechanical Turk to evaluate how they felt about various subjects, such as camping, health care, architecture, taxidermy, crossword puzzles, and Japan in order to measure their attitude towards mostly independent stimuli. Then, they presented the participants with information about a new product: a microwave oven. This microwave oven does not exist, but the participants didn't know this, and were given three positive and three negative fake reviews. People who reacted positively to the subjects on the dispositional attitude measurement also tended to react positively to the microwave oven, and those who reacted negatively tended to react negatively to it. Researchers concluded that *"some people tend to like things, whereas others tend to dislike things, and a more thorough understanding of this tendency will lead to a more thorough understanding of the psychology of attitudes."* (Hepler and Albarracín 2013)

    a. What are the cases?

    b. What is (are) the response variable(s) in this study?

    c. What is (are) the explanatory variable(s) in this study?

    d. Does the study employ random sampling? Explain your reasoning.

    e. Is this an observational study or an experiment? Explain your reasoning.

    f. Can we establish a causal link between the explanatory and response variables?

    g. Can the results of the study be generalized to the population at large?

16. **Reading the paper.** Below are excerpts from two articles published in the *NY Times*:

    a. An excerpt from an article titled *Risks: Smokers Found More Prone to Dementia* is below. Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning. (Rabin 2010)

    > "Researchers analyzed data from 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50-60 years old. 23 years later, about 25% of the group had dementia, including 1,136 with Alzheimer's disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up with increased smoking; 44% for one to two packs a day; and twice the risk for more than two packs."

    b. An excerpt from an article titled *The School Bully Is Sleepy* is below. A friend of yours who read the article says, *"The study shows that sleep disorders lead to bullying in school children."* Is this statement justified? If not, how best can you describe the conclusion that can be drawn from this study? (Parker-Pope 2011)

    > "The University of Michigan study, collected survey data from parents on each child's sleep habits and asked both parents and teachers to assess behavioral concerns. About a third of the students studied were identified by parents or teachers as having problems with disruptive behavior or bullying. The researchers found that children who had behavioral issues and those who were identified as bullies were twice as likely to have shown symptoms of sleep disorders."

17. **Sampling strategies.** A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey. Various research strategies for collecting data are described below. In each, name the sampling method proposed and any bias you might expect.

a. They randomly sample 40 students from the study's population, give them the survey, ask them to fill it out, and bring it back the next day.

b. They give out the survey only to their friends, making sure each one of them fills it out.

c. They post a link to an online survey on Facebook and ask their friends to fill it out.

d. They randomly sample 5 classes and asks a random sample of students from those classes to fill out the survey.

18. **Family size.** Suppose we want to estimate household size, where a *"household"* is defined as people living together in the same dwelling, and sharing living accommodations. If we select students at random at an elementary school and ask them what their family size is, will this be a good measure of household size? Or will our average be biased? If so, will it overestimate or underestimate the true value?

19. **Light and exam performance.** A study is designed to test the effect of light level on exam performance of students. The researcher believes that light levels might have different effects on people who wear glasses and people who don't, so they want to make sure both groups of people are equally represented in each treatment. The treatments are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps).

    a. What is the response variable?

    b. What is the explanatory variable? What are its levels?

    c. What is the blocking variable? What are its levels?

20. **Vitamin supplements.** To assess the effectiveness of taking large doses of vitamin C in reducing the duration of the common cold, researchers recruited 400 healthy volunteers from staff and students at a university. A quarter of the patients were assigned a placebo, and the rest were evenly divided between 1g Vitamin C, 3g Vitamin C, or 3g Vitamin C plus additives to be taken at onset of a cold for the following two days. All tablets had identical appearance and packaging. The nurses who handed the prescribed pills to the patients knew which patient received which treatment, but the researchers assessing the patients when they were sick did not. No statistically discernible differences were observed in any measure of cold duration or severity between the four groups, and the placebo group had the shortest duration of symptoms. (Audera et al. 2001)

    a. Was this an experiment or an observational study? Why?

    b. What are the explanatory and response variables in this study?

    c. Were the patients blinded to their treatment?

    d. Was this study double-blind?

    e. Participants are ultimately able to choose whether to use the pills prescribed to them. We might expect that not all of them will adhere and take their pills. Does this introduce a confounding variable to the study? Explain your reasoning.

21. **Light, noise, and exam performance.** A study is designed to test the effect of light level and noise level on exam performance of students. The researcher believes that light and noise levels might have different effects on people who wear glasses and people who don't, so they want to make sure both groups of people are equally represented in each treatment. The light treatments considered are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). The noise treatments considered are no noise, construction noise, and human chatter noise.

    a. What type of study is this?

    b. How many factors are considered in this study? Identify them, and describe their levels.

    c. What is the role of the wearing glasses variable in this study?

22. **Music and learning.** You would like to conduct an experiment in class to see if students learn better if they study without any music, with music that has no lyrics (instrumental), or with music that has lyrics. Briefly outline a design for this study.

23. **Soda preference.** You would like to conduct an experiment in class to see if your classmates prefer the taste of regular Coke or Diet Coke. Briefly outline a design for this study.

24. **Exercise and mental health.** A researcher is interested in the effects of exercise on mental health and they propose the following study: use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41- 55 year-olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.

    a. What type of study is this?

    b. What are the treatment and control groups in this study?

    c. Does this study make use of blocking? If so, what is the blocking variable?

    d. Does this study make use of blinding?

    e. Comment on whether the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether the conclusions can be generalized to the population at large.

    f. Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?

25. **Chia seeds and weight loss.** Chia Pets – those terra-cotta figurines that sprout fuzzy green hair – made the chia plant a household name. But chia has since gained a reputation as a diet supplement. In one 2009 study, 38 men and 38 women were recruited and and divided each randomly into two groups: treatment or control. One group was given 25 grams of chia seeds twice a day, and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no statistically discernible difference between the groups in appetite or weight loss. (Nieman et al. 2009)

    a. What type of study is this?

    b. What are the experimental and control treatments in this study?

    c. Has blocking been used in this study? If so, what is the blocking variable?

    d. Has blinding been used in this study?

    e. Comment on whether we can make a causal statement, and indicate whether we can generalize the conclusion to the population at large.

26. **City council survey.** A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. For each part below, identify the sampling methods described, and describe the statistical pros and cons of the method in the city's context.

    a. Randomly sample 200 households from the city.

    b. Divide the city into 20 neighborhoods, and then sample 10 households from each neighborhood.

    c. Divide the city into 20 neighborhoods, randomly sample 3 neighborhoods, and then sample all households from those 3 neighborhoods.

    d. Divide the city into 20 neighborhoods, randomly sample 8 neighborhoods, and then randomly sample 50 households from those neighborhoods.

    e. Sample the 200 households closest to the city council offices.

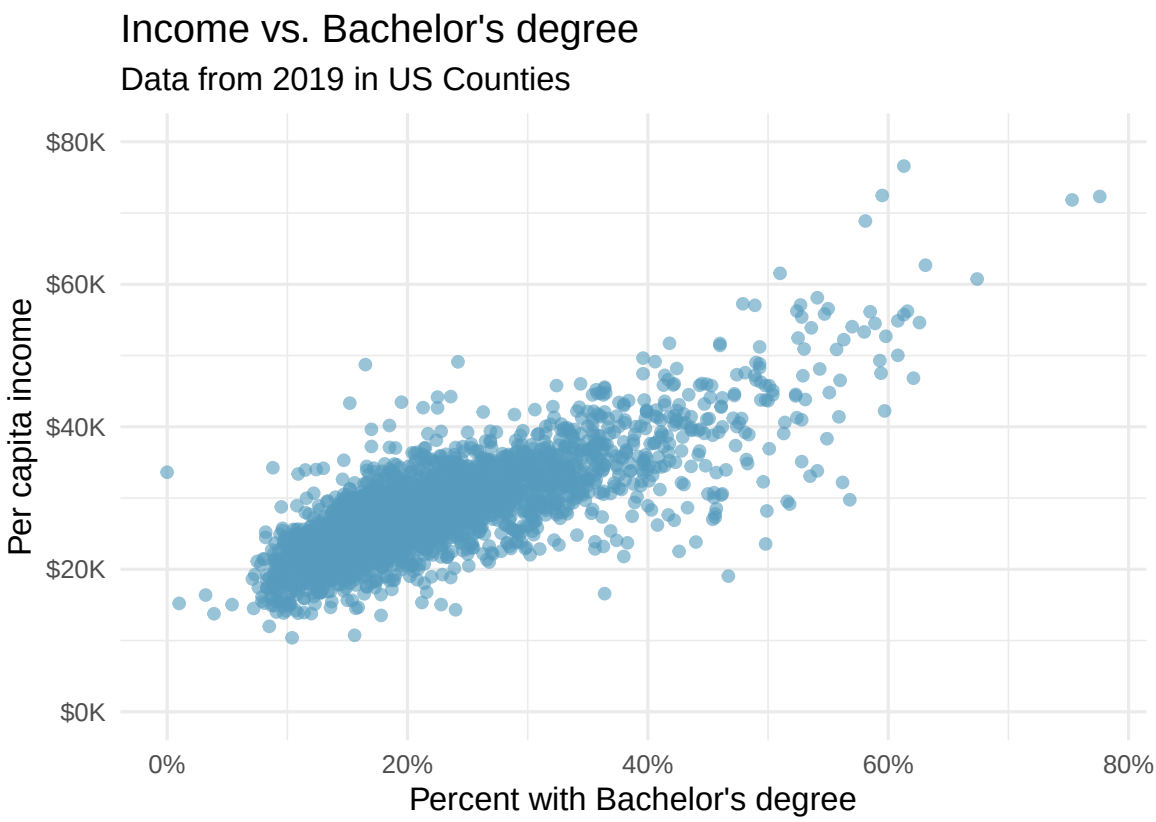
27. **Flawed reasoning.** Identify the flaw(s) in reasoning in the following scenarios. Explain what the individuals in the study should have done differently if they wanted to make such strong conclusions.

    a. Students at an elementary school are given a questionnaire that they are asked to return after their parents have completed it. One of the questions asked is, *"Do you find that your work schedule makes it difficult for you to spend time with your kids after school?"* Of the parents who replied, 85% said *"no"*. Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.

    b. A survey is conducted on a simple random sample of 1,000 women who recently gave birth, asking them about whether they smoked during pregnancy. A follow-up survey asking if the children have respiratory problems is conducted 3 years later. However, only 567 of these women are reached at the same address. The researcher reports that these 567 women are representative of all mothers.

    c. An orthopedist administers a questionnaire to 30 of his patients who do not have any joint problems and finds that 20 of them regularly go running. He concludes that running decreases the risk of joint problems.

28. **Income and education in US counties.** The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,142 counties in the US in 2019.



    a. What are the explanatory and response variables?

    b. Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.

c. Can we conclude that having a bachelor's degree increases one's income?

29. **Eat well, feel better.** In a public health study on the effects of consumption of fruits and vegetables on psychological well-being in young adults, participants were randomly assigned to three groups: (1) diet-as-usual, (2) an ecological momentary intervention involving text message reminders to increase their fruits and vegetable consumption plus a voucher to purchase them, or (3) a fruit and vegetable intervention in which participants were given two additional daily servings of fresh fruits and vegetables to consume on top of their normal diet. Participants were asked to take a nightly survey on their smartphones. Participants were student volunteers at the University of Otago, New Zealand. At the end of the 14-day study, only participants in the third group showed improvements to their psychological well-being across the 14-days relative to the other groups. (Conner et al. 2017)

    a. What type of study is this?

    b. Identify the explanatory and response variables.

    c. Comment on whether the results of the study can be generalized to the population.

    d. Comment on whether the results of the study can be used to establish causal relationships.

    e. A newspaper article reporting on the study states, "The results of this study provide proof that giving young adults fresh fruits and vegetables to eat can have psychological benefits, even over a brief period of time." How would you suggest revising this statement so that it can be supported by the study?

30. **Screens, teens, and psychological well-being.** In a study of three nationally representative large-scale datasets from Ireland, the United States, and the United Kingdom (n = 17,247), teenagers between the ages of 12 to 15 were asked to keep a diary of their screen time and answer questions about how they felt or acted. The answers to these questions were then used to compute a psychological well-being score. Additional data were collected and included in the analysis, such as each child's sex and age, and on the mother's education, ethnicity, psychological distress, and employment. The study concluded that there is little clear-cut evidence that screen time decreases adolescent well-being. (Orben and Baukney-Przybylski 2018)

    a. What type of study is this?

    b. Identify the explanatory variables.

    c. Identify the response variable.

    d. Comment on whether the results of the study can be generalized to the population, and why.

    e. Comment on whether the results of the study can be used to establish causal relationships.

Datasets: `cia_factbook` (openintro) | `county_complete` (openintro)

Chapter 3

Exploring Categorical Data

This chapter focuses on **categorical data** — variables whose values are categories or labels rather than numbers. The distinct values a categorical variable can take are called its **levels**. For example, *homeownership status* is a single categorical variable with three levels (rent, mortgage, own) — it is **one** variable, not three, even though it has three levels. Keeping the variable separate from its levels avoids a common miscount. We begin with visualizations for a single categorical variable (bar charts, pie charts, waffle charts), then move to tools for exploring the relationship between two categorical variables (contingency tables and stacked bar charts). We close with principles for creating effective, honest data visualizations.

3.1 Visualizing one categorical variable

3.1.1 Bar charts

A **bar chart** is the most common way to display the distribution of a single categorical variable. Each level (category) gets its own bar, and the height (or length) of the bar represents either the count or the proportion of observations in that level.

Consider a dataset of 10,000 loans from a lending platform. Each borrower's homeownership status is recorded as `rent`, `mortgage`, or `own`. The frequency table below summarizes the data:

Homeownership	Count
Rent	3,858
Mortgage	4,789
Own	1,353
Total	**10,000**

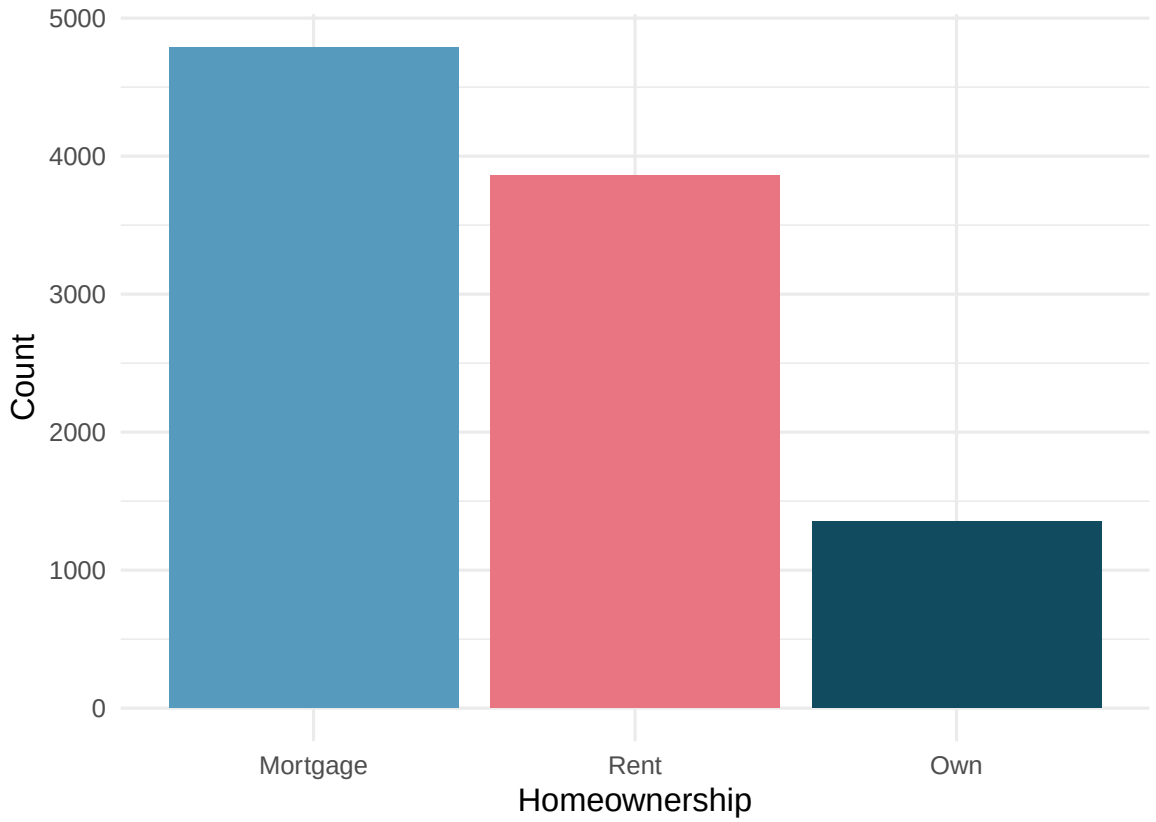


Figure 3.1: Bar chart showing counts of homeownership status. Mortgage is the tallest bar, followed by rent, then own. Open in StatLens

A bar chart of proportions conveys the same information but rescales the vertical axis so the bars represent the fraction of observations in each category. In the homeownership example, about 47.9% of borrowers have a mortgage, 38.6% rent, and 13.5% own their home outright.

To find a **proportion**, divide the number of observations in a category by the total number of observations. We write a sample proportion as $\hat{p}$ (read "p-hat"). For instance, the proportion of borrowers who rent is

$$\hat{p} = \frac{\text{number who rent}}{\text{total number of borrowers}} = 0.386, \text{ or } 38.6\%.$$

We will use $\hat{p}$ throughout the course, so it is worth getting comfortable computing a proportion now.

In the homeownership bar chart, which category is most common? Which is least common? About what fraction of borrowers rent?

Show answer

Mortgage is the most common category and own is the least common. About 38.6% of borrowers rent their home.

3.1.2 Pie charts

A **pie chart** divides a circle into slices, where each slice represents a category and the size of the slice corresponds to the proportion of observations in that category.

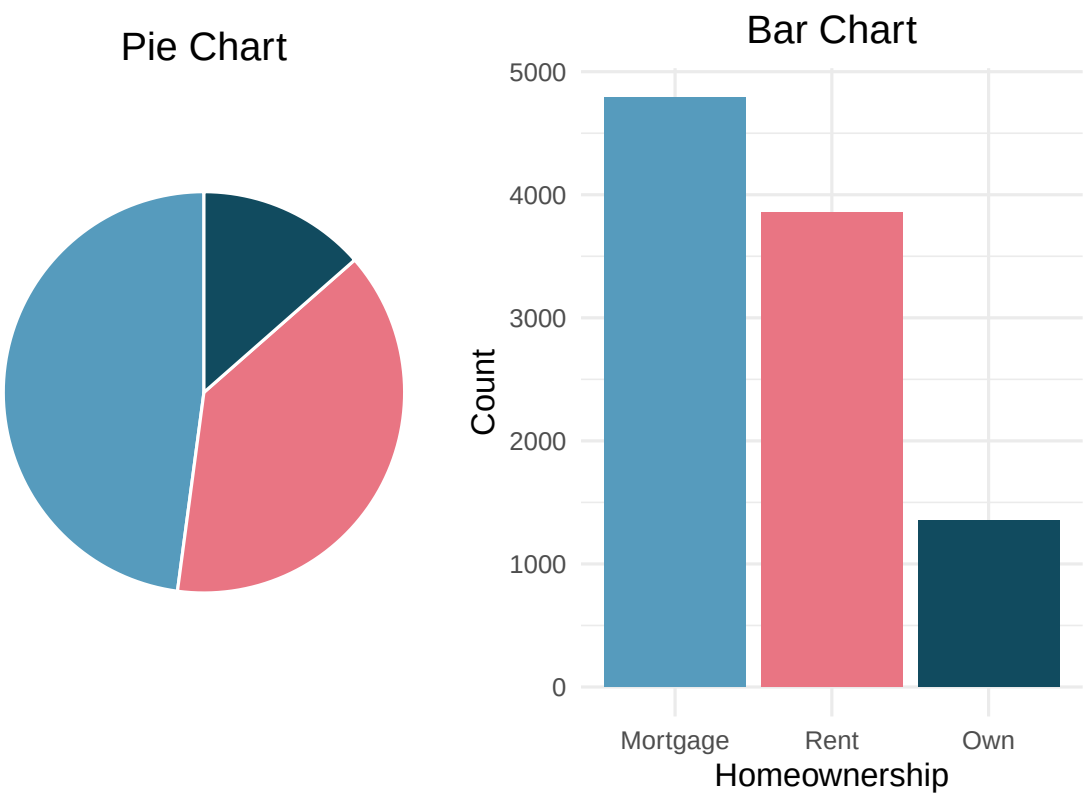


Figure 3.2: A pie chart and bar chart displaying the same homeownership data side by side. Comparing mortgage and rent is much easier in the bar chart. Open in StatLens

Pie charts can give a quick high-level overview, especially when a variable has only a few levels or when each level represents a simple fraction (one-half, one-quarter, etc.). However, bar charts are almost always easier to read because our eyes are better at comparing lengths than angles or areas.

Bar charts vs. pie charts. When a categorical variable has many levels, bar charts are far superior to pie charts for making comparisons. Reserve pie charts for situations where the variable has very few categories and the goal is to show how a whole breaks into parts.

3.1.3 Waffle charts

A **waffle chart** is another way to visualize proportions. It uses a grid (often 10 by 10 = 100 squares) where each square represents 1% of the data. Squares are shaded according to the categories. Like pie charts, waffle charts work best when the number of categories is small, but they can make it easier to compare proportions that do not correspond to simple fractions.

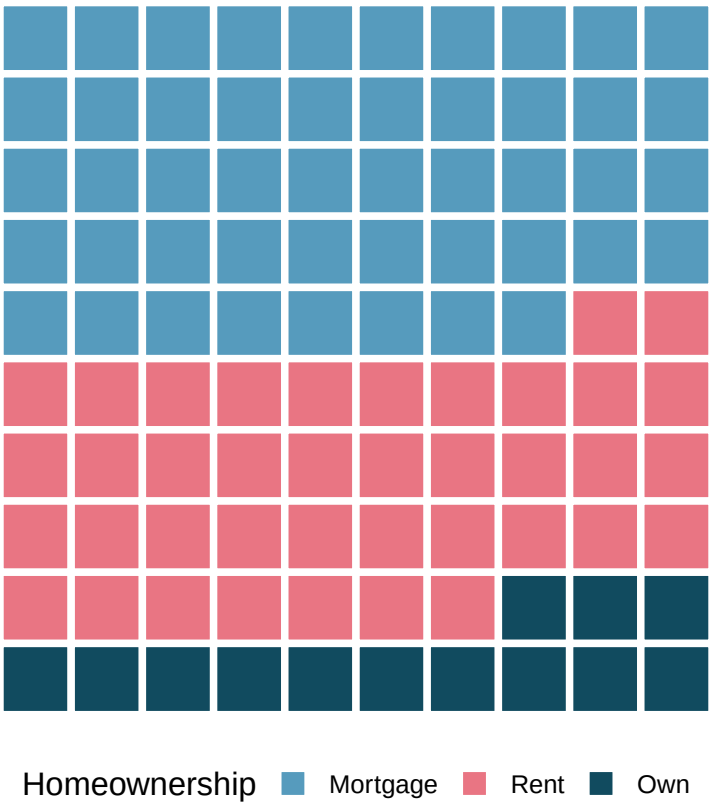


Figure 3.3: A waffle chart of homeownership status. Each square represents 1% of borrowers. Open in StatLens

Why might a waffle chart be easier to read than a pie chart when comparing two categories with similar proportions (say, 38% and 42%)?

Show answer

In a waffle chart, you can count the number of shaded squares for each category, making it straightforward to see the difference. In a pie chart, the angle difference between 38% and 42% is very subtle and hard to judge visually.

See it in action. Open the Comparing Charts for Categorical Data activity to switch between bar, pie, and waffle displays on Brexit survey data and see which makes it easiest to compare five response categories.

3.2 Two categorical variables

When we want to explore the relationship between two categorical variables, we organize the data in a **contingency table** and display it with stacked or grouped bar charts.

3.2.1 Contingency tables

A **contingency table** summarizes data for two categorical variables. The rows represent the levels of one variable, the columns represent the levels of the other, and each cell contains the count of observations in that combination. **Row totals** and **column totals** give the marginal distributions.

Consider a dataset of 10,000 loans. Each loan has a homeownership status (`rent`, `mortgage`, or `own`) and an application type (`joint` or `individual`). The contingency table below summarizes these two variables:

	Rent	Mortgage	Own	Total
Joint	362	950	183	1,495
Individual	3,496	3,839	1,170	8,505
Total	**3,858**	**4,789**	**1,353**	**10,000**

Each cell represents a count. For example, 3,496 loans were made by individuals who rent their home. The row totals tell us that 1,495 loans were joint applications and 8,505 were individual applications. The column totals tell us the overall distribution of homeownership.

What does the value 950 in the table represent? What does the column total 4,789 represent?

Show answer

The value 950 represents the number of loans that were joint applications from borrowers with a mortgage. The column total 4,789 represents the total number of borrowers (both joint and individual) who have a mortgage.

3.2.2 Row and column proportions

Raw counts can be hard to compare when group sizes differ. Instead, we can compute **row proportions** or **column proportions** to see conditional relationships.

Row proportions are computed by dividing each count by its row total. They show how a variable is distributed *within each level* of the row variable.

Column proportions are computed by dividing each count by its column total. They show how a variable is distributed *within each level* of the column variable.

Both row and column proportions are examples of **conditional proportions** — the proportion of observations in one category, conditional on (given) the level of the other variable.

Row proportions (each count divided by its row total):

	Rent	Mortgage	Own	Total
Joint	0.242	0.635	0.122	1.000
Individual	0.411	0.451	0.138	1.000

For example, $3{,}496/8{,}505 = 0.411$, meaning 41.1% of individual applicants rent their home.

Column proportions (each count divided by its column total):

	Rent	Mortgage	Own
Joint	0.094	0.198	0.135
Individual	0.906	0.802	0.865
Total	1.000	1.000	1.000

The value 0.906 tells us that 90.6% of renters applied individually. This rate is higher than for borrowers with mortgages (80.2%) or who own their home (86.5%). Because these rates vary across

the levels of homeownership, there is evidence that application type and homeownership may be **associated**.

What does 0.451 represent in the row proportions table? What does 0.802 represent in the column proportions table?

Show answer

0.451 represents the proportion of individual applicants who have a mortgage. 0.802 represents the fraction of applicants with mortgages who applied as individuals.

Data scientists use statistics to build email spam filters. One characteristic examined is email format (HTML or plain text). A contingency table of email format and spam status is shown below:

	HTML	Text	Total
Not spam	2,568	986	3,554
Spam	158	209	367
Total	**2,726**	**1,195**	**3,921**

Which would be more useful for classifying email: row or column proportions?

A data scientist would want to know how the proportion of spam changes depending on email format. This corresponds to **column proportions**: $209/1{,}195 = 17.5\%$ of plain text emails are spam, compared to $158/2{,}726 = 5.8\%$ of HTML emails. Plain text emails are more likely to be spam. Row or column proportions are not equivalent — before computing, consider which variable you want to "condition on" (usually the explanatory variable).

In the spam example, which variable would you consider the explanatory variable and which the response?

Show answer

Email format (HTML vs. text) is the explanatory variable — it is a characteristic of the email that might help explain or predict whether the email is spam. Spam status (spam vs. not spam) is the response variable.

3.2.3 Bar charts for two categorical variables

Bar charts can be extended to display the relationship between two categorical variables. There are three common variants:

A **stacked bar chart** stacks the bars for the second variable on top of each other within each level of the first variable. It shows both the total count and the breakdown by the second variable.

A **100% stacked bar chart** (also called a standardized or filled bar chart) is like a stacked bar chart, but the bars are scaled to the same height (100%). It shows the *proportion* of each category of the second variable within each level of the first variable.

A **side-by-side bar chart** (also called a grouped or dodged bar chart) places bars for the second variable next to each other within each level of the first variable. It makes it easy to compare counts across groups.

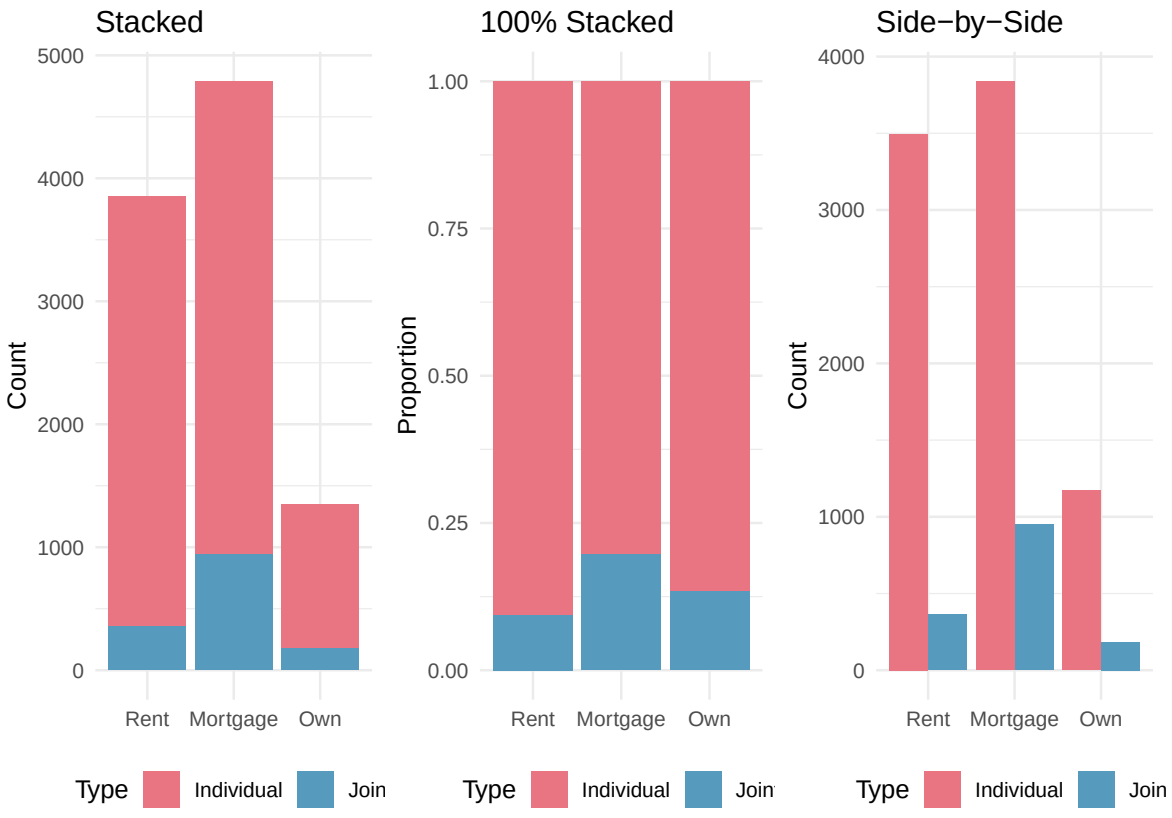


Figure 3.4: Three bar charts showing homeownership and application type: stacked, 100% stacked, and side-by-side. Open in StatLens

When is each type of bar chart most useful?

- The **stacked bar chart** is most useful when you want to see both the total count for each group and the breakdown by the second variable.
- The **100% stacked bar chart** is helpful when comparing proportions across groups, especially when the groups have very different sizes. However, it sacrifices information about the total count.
- The **side-by-side bar chart** makes it easy to compare individual counts directly. It works well for showing both variables without implying one is the explanatory variable. However, it requires more horizontal space.

Detecting association in a 100% stacked bar chart. If the proportional breakdown of the response variable is the *same* across all levels of the explanatory variable, then the variables are **not** associated — the colored segments will line up at the same heights. If the breakdown *differs* across levels, the variables are associated.

In the 100% stacked bar chart above, how can you tell whether homeownership and application type are associated?

Show answer

If the variables were not associated, the division between joint and individual would occur at the same height in every bar. Since the bars are divided at different heights — mortgage holders have a higher

proportion of joint applications than renters or owners — the variables appear to be associated.

See it in action. Open the Exploring Two-Way Tables activity to walk through counts, row / column / joint proportions, and three bar-chart types on a diabetes clinical trial's contingency table.

3.3 Effective data visualization

Graphs can powerfully communicate ideas directly and quickly. However, there are times when a visualization conveys a message that is inaccurate or misleading. This section provides guiding principles for creating clear, honest, and effective visualizations.

3.3.1 Keep it simple

Colors should be used purposefully — to group items or differentiate levels in meaningful ways. Colors used only for decoration can be distracting.

Consider a company that has summarized its costs into five categories. A three-dimensional pie chart makes it hard to compare the sizes of the slices, and the unnecessary 3D effect distorts the proportions. A simple bar chart conveys the same information much more clearly.

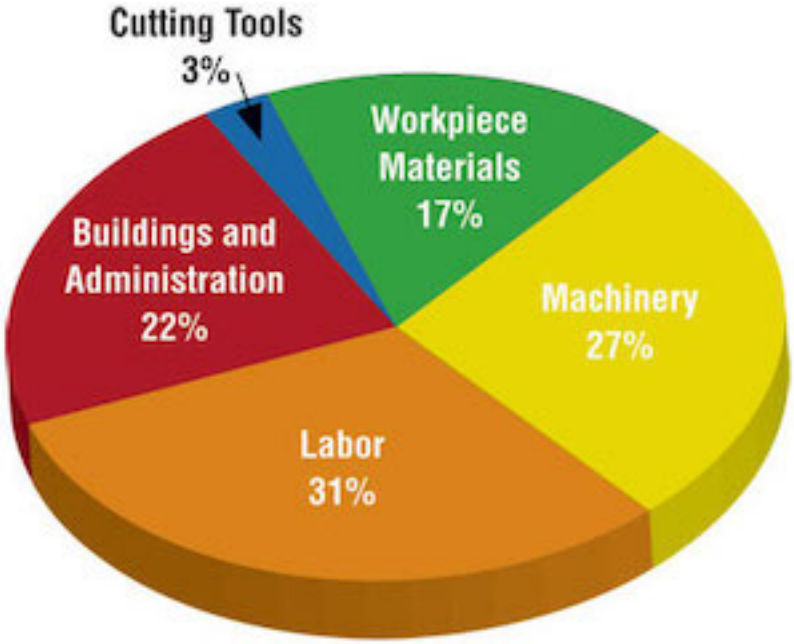


Figure 3.5: A 3D pie chart showing five cost categories. The 3D perspective distorts the proportions — for instance, "Labor" at 31% looks disproportionately large because it is in the foreground.

A simple bar chart of the same data makes the comparisons effortless:

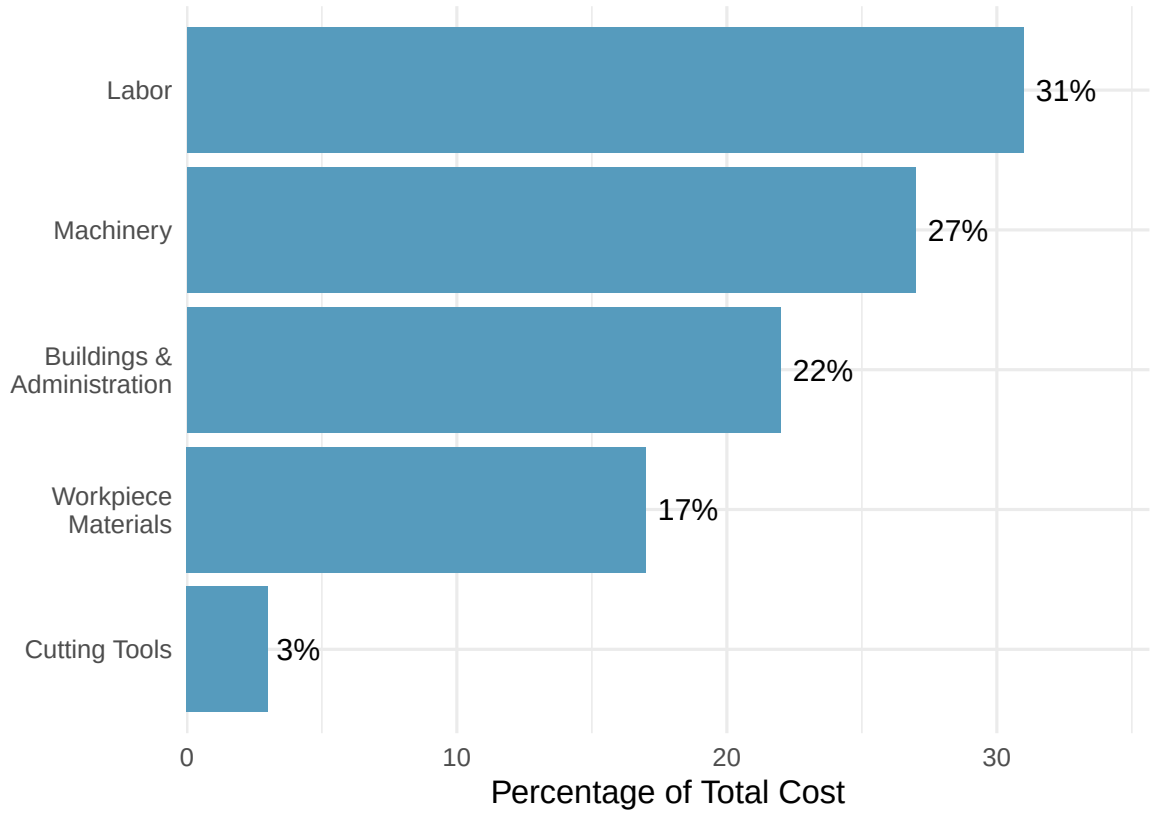


Figure 3.6: A bar chart of the same cost data. The relative sizes of each category are immediately clear.

3.3.2 Use color to draw attention

An important principle is to use color **to draw attention** to the most important part of your graphic. Default rainbow coloring often adds no meaning and can confuse readers.

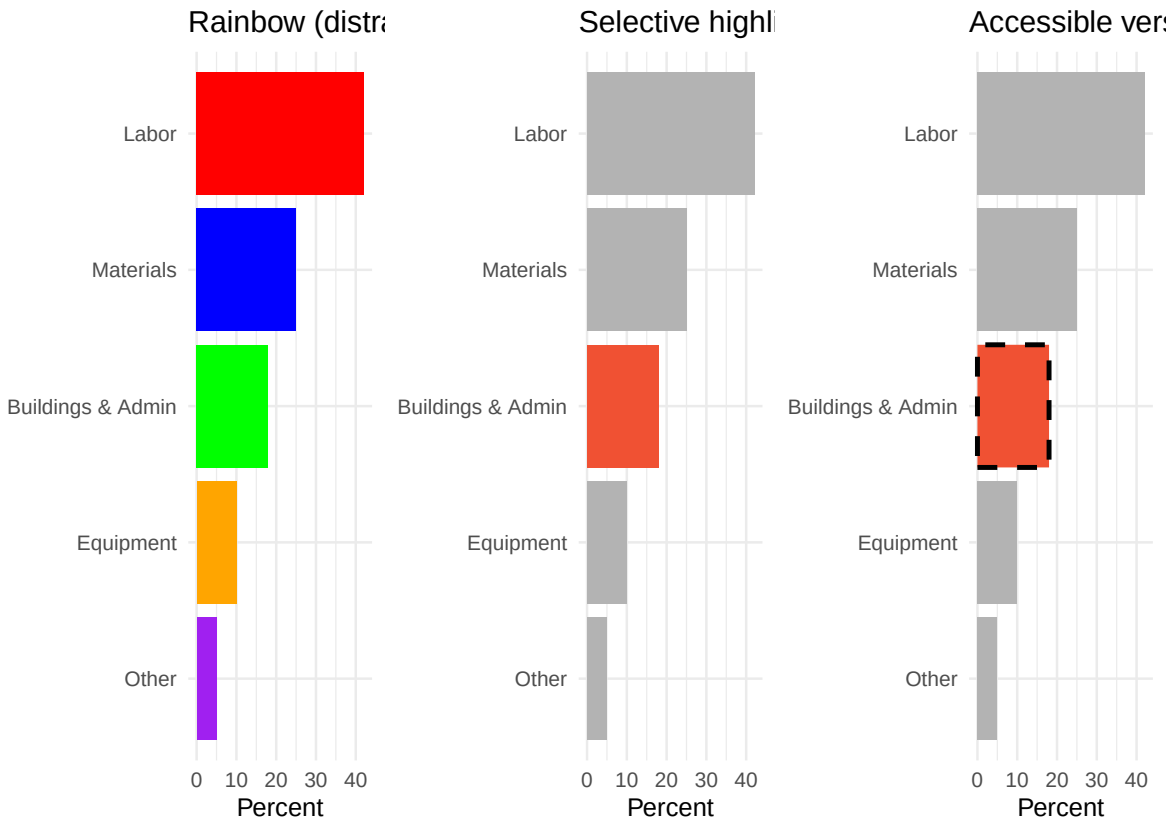


Figure 3.7: Three versions of the same bar chart. The first uses distracting rainbow colors. The second highlights one important bar in red against gray bars. The third adds both color and a dashed border for accessibility.

Accessibility matters. Not everyone perceives color the same way. When using color to draw attention, also consider adding a secondary visual cue (such as a pattern, line style, or border) so that the highlighted feature is distinguishable even for people with color vision deficiency.

3.3.3 Tell a story

Effective graphs often include **annotations** — text, lines, or labels that provide context beyond the raw data. For example, adding a note that "July is the start of the fiscal year" to a time series of monthly hiring can explain why there is a spike every July.

3.3.4 Order matters

The arrangement of categories in a bar chart can help or hinder understanding. Alphabetical order is rarely the most informative choice. Consider ordering bars by frequency (tallest to shortest) or by a meaningful sequence (such as the order the options appeared in a survey question).

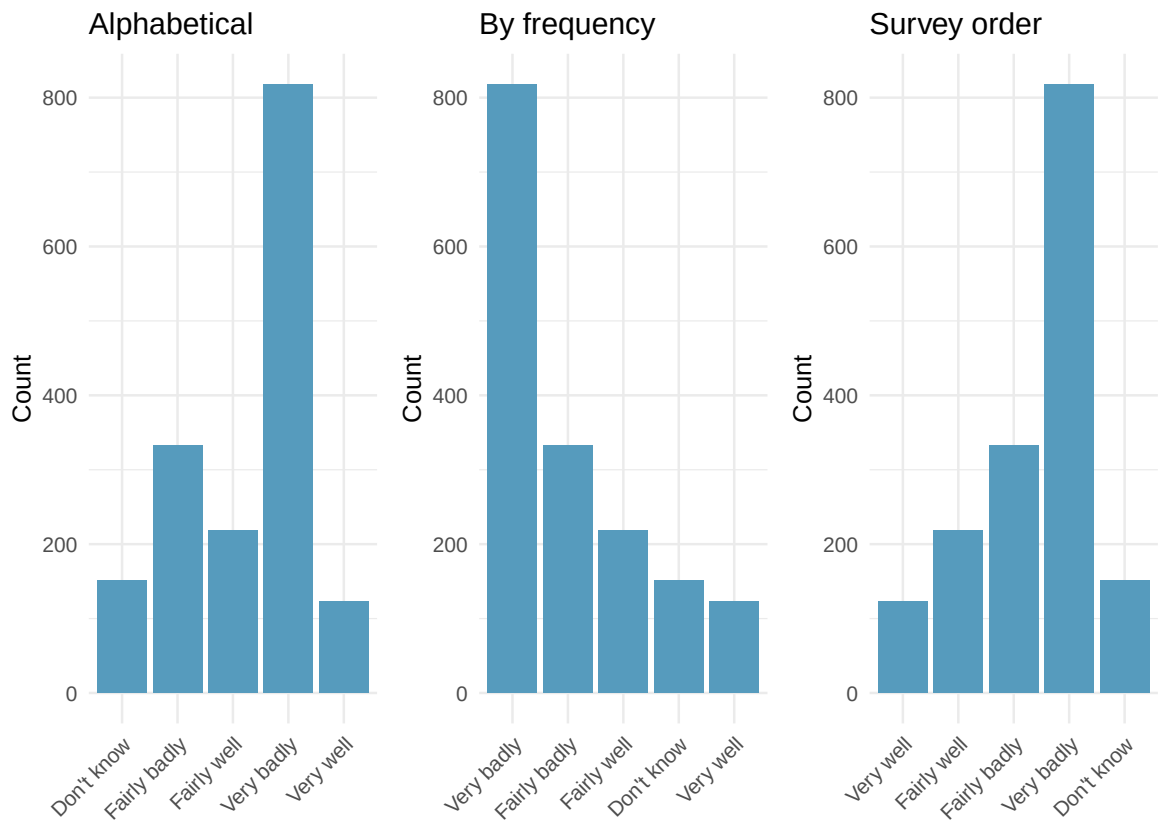


Figure 3.8: Three versions of a bar chart showing responses from 1,643 people in a 2019 YouGov survey about how well the UK government was handling Brexit, ordered alphabetically, by frequency, and by the original survey order. Open in StatLens

3.3.5 Make labels readable

When category labels are long, consider using horizontal bars (where labels appear along the vertical axis with plenty of space) rather than vertical bars (where labels must be angled or truncated).

3.3.6 Pick a purpose

Every graphical decision should be made with a purpose. Before choosing a chart type, ask:

- What is the key comparison I want the reader to make?
- Which chart type makes that comparison easiest?
- Am I showing proportions, counts, distributions, or relationships?

A survey asks residents of five regions about their opinion on a policy, with responses ranging from "Very well" to "Very badly." You want to compare how opinions differ across regions. What visualization would you choose?

A 100% stacked bar chart would be useful if the main comparison is about proportions within each region. A side-by-side bar chart would work if comparing raw counts across regions matters. A faceted bar chart — one panel per region — would allow the most detailed comparison. The best choice depends on which comparison is most important to the story you are telling.

3.3.7 Select meaningful colors

Default or rainbow color schemes are not always the best choice. Perceptually uniform color scales (such as the viridis or cividis scales) are designed to be distinguishable by people with a variety of color vision abilities and to be readable when printed in grayscale. For ordinal data, a sequential color scale (light to dark) communicates the ordering effectively.

Summary of data visualization principles:

1. **Keep it simple** — avoid unnecessary 3D effects, decorative colors, and chart junk
2. **Use color with purpose** — to group, differentiate, or draw attention
3. **Include context** — titles, axis labels, annotations, and data sources
4. **Choose appropriate ordering** — by frequency, natural order, or meaningful sequence
5. **Make labels readable** — flip axes if labels are long; avoid angled text
6. **Consider your audience** — use accessible color schemes and secondary visual cues
7. **Match the chart to the question** — each chart type has strengths; choose deliberately

3.4 Chapter review

3.4.1 Summary

This chapter introduced tools for exploring categorical data. For a single categorical variable, **bar charts** are the primary visualization, with **pie charts** reserved for simple cases and **waffle charts** useful for estimating percentages. For two categorical variables, **contingency tables** organize counts, **row and column proportions** reveal conditional relationships, and **stacked**, **100% stacked**, and **side-by-side bar charts** provide visual comparisons. When the proportional breakdown of one variable differs across levels of the other, the variables are **associated**. Effective visualizations follow principles of simplicity, purposeful use of color, meaningful ordering, and accessibility.

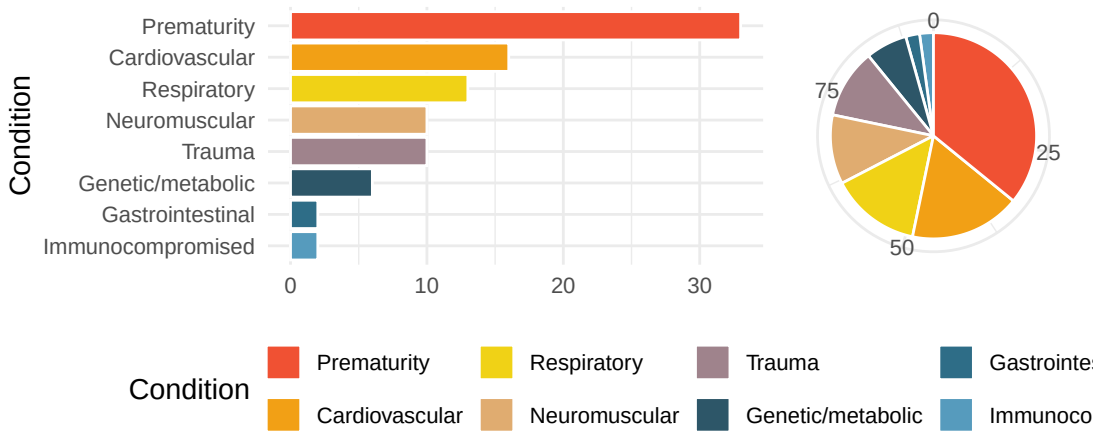
3.4.2 Key terms

- Bar chart, pie chart, waffle chart
- Frequency table, proportion
- Contingency table, row totals, column totals
- Row proportions, column proportions, conditional proportions
- Stacked bar chart, 100% stacked bar chart, side-by-side bar chart
- Association (between categorical variables)
- Explanatory variable, response variable

3.5 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Antibiotic use in children.** The bar plot and the pie chart below show the distribution of pre-existing medical conditions of children involved in a study on the optimal duration of antibiotic use in treatment of tracheitis, which is an upper respiratory infection.



a. What features are apparent in the bar plot but not in the pie chart?

b. What features are apparent in the pie chart but not in the bar plot?

c. Which graph would you prefer to use for displaying these categorical data?

2. **Views on immigration.** Nine-hundred and ten (910) randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.

Response	Conservative	Liberal	Moderate	Total
Apply for citizenship	57	101	120	278
Guest worker	121	28	113	262
Leave the country	179	45	126	350
Not sure	15	1	4	20
Total	372	175	363	910

a. What percent of these voters identify themselves as conservatives?

b. What percent of these voters are in favor of the citizenship option?

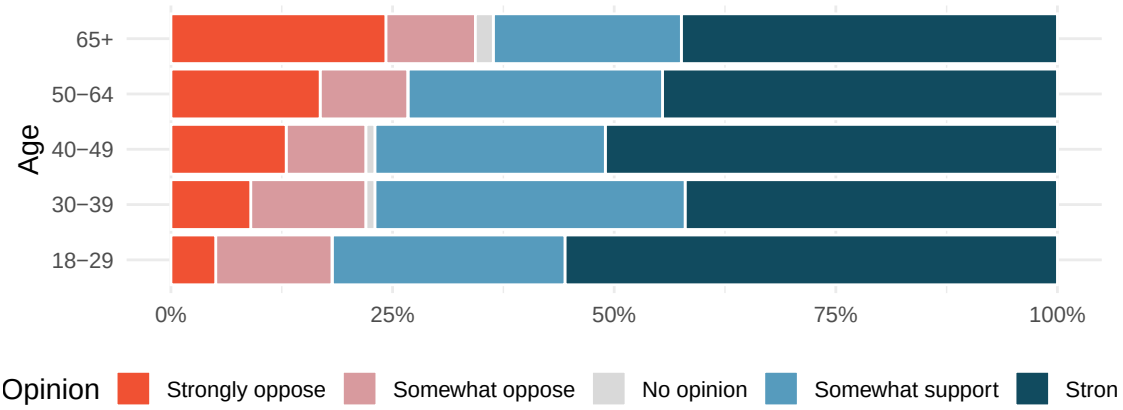
c. What percent of these voters identify themselves as conservatives *and* are in favor of the citizenship option?

d. What percent of these voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

e. Do political ideology and views on immigration appear to be associated? Explain your reasoning.

f. Conjecture other variables that might explain the potential relationship between these two variables.

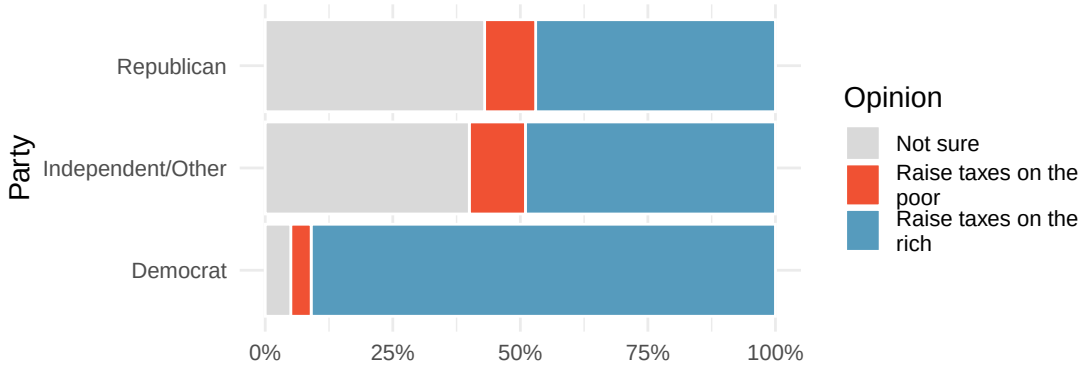
3. **Black Lives Matter.** A Washington Post-Schar School poll conducted in the United States in June 2020, among a random national sample of 1,006 adults, asked respondents whether they support or oppose protests following George Floyd's killing that have taken place in cities across the US. The survey also collected information on the age of the respondents. (Washington Post 2020) The results are summarized in the stacked bar plot below.



a. Based on the stacked bar plot, do views on the protests and age appear to be associated? Explain your reasoning.

b. Conjecture other possible variables that might explain the potential association between these two variables.

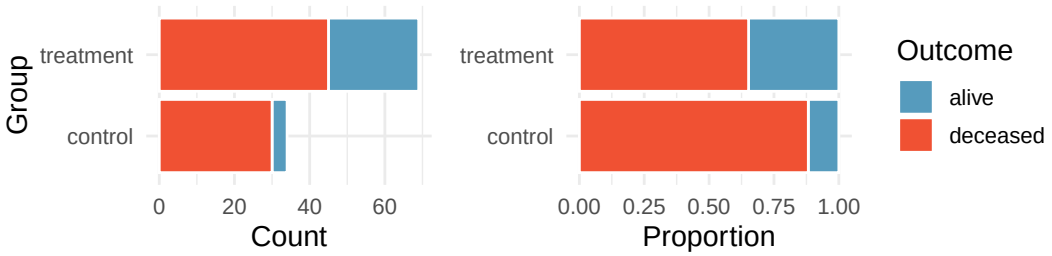
4. **Raise taxes.** A random sample of registered voters nationally were asked whether they think it's better to raise taxes on the rich or raise taxes on the poor. The survey also collected information on the political party affiliation of the respondents. (Polling 2015)



a. Based on the stacked bar plot shown above, do views on raising taxes and political affiliation appear to be associated? Explain your reasoning.

b. Conjecture other possible variables that might explain the potential association between these two variables.

5. **Heart transplant data display.** The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was officially designated a heart transplant candidate, meaning that they were gravely ill and might benefit from a new heart. Patients were randomly assigned into treatment and control groups. Patients in the treatment group received a transplant, and those in the control group did not. The visualizations below display two different versions of the study results. (Turnbull, Brown, and Hu 1974)

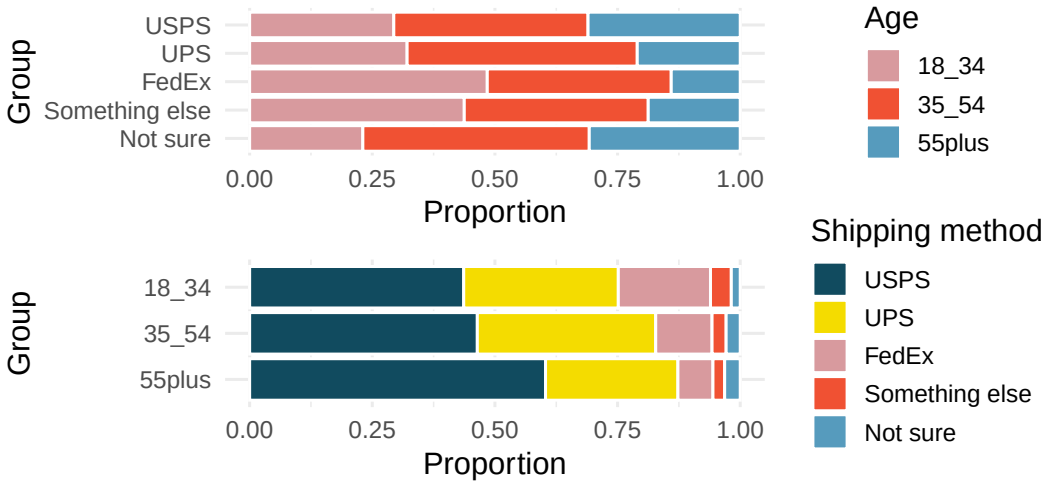


a. Provide one aspect of the two group comparison that is easier to see from the stacked bar plot (left)?

b. Provide one aspect of the two group comparison that is easeir to see from the standardized bar plot (right)?

c. For the Heart Transplant Study which of those aspects would be more important to display? That is, which bar plot would be better as a data visualization?

6. **Shipping holiday gifts data display.** A local news survey asked 500 randomly sampled Los Angeles residents which shipping carrier they prefer to use for shipping holiday gifts. The bar plots below show the distribution of responses by age group as well as distribution of responses by shipping method.



a. Which graph (top or bottom) would you use to understand the shipping choices of people of different ages? Explain.

b. Which graph (top or bottom) would you use to understand the age distribution across different types of shipping choices? Explain.

c. A new shipping company would like to market to people over the age of 55. Who will be their biggest competitor? Explain.

d. FedEx would like to reach out to grow their market share so as to balance the age demographics of FedEx users. To what age group should FedEx market?

7. **On-time arrivals.** Consider all of the flights out of New York City in 2013 that flew into Puerto Rico (BQN), Los Angeles (LAX), or San Francisco (SFO) on the following two airlines: JetBlue (B6) or United Airlines (UA). Below are the tabulated counts for the number of flights `delayed` and `on time` for each airline into each city.

dest	carrier	status	count
BQN	B6	delayed	271
BQN	B6	on time	322
BQN	UA	delayed	144
BQN	UA	on time	151
LAX	B6	delayed	670
LAX	B6	on time	999
LAX	UA	delayed	2368
LAX	UA	on time	3402
SFO	B6	delayed	405
SFO	B6	on time	615
SFO	UA	delayed	2694
SFO	UA	on time	4034

a. What percent of all JetBlue flights were delayed? What percent of all United Airlines flights were delayed? (Note, the overall delay proportions are typically what would be reported and associated with an airline.)

b. For each of the three airports, find the percent of delayed flights for each of JetBlue and United (you should have 6 numbers).

c. United has a higher proportion of delayed flights for each of the three cities, yet JetBlue has a higher proportion of delayed flights overall. Explain, using the data counts provided, how the seeming paradox could happen.

8. **US House of Representatives.** The US House of Representatives is dominated by two political parties: Democrats and Republicans. Democrats are thought to be the more liberal party and Republicans are considered to be the more conservative party. However, within each party there is an internal spectrum of liberal to conservative. For example, conservative Democrats and liberal Republicans would be labeled moderate. Consider an election where the only change in membership is that the most conservative Democrats are replaced by a set of liberal Republicans who are more liberal than the incumbent Republicans but more conservative than the Democrats they replaced.

a. After the election, is the Democratic wing of the House more conservative or more liberal? Explain.

b. After the election, is the Republican wing of the House more conservative or more liberal? Explain.

c. After the election, is the overall House membership more conservative or more liberal? Explain.

d. In what settings would you report the outcome of the change in House membership to be more conservative? And in what settings would you report the outcome of the change in House membership to be more liberal?

Datasets: `antibiotics` (openintro) | `immigration` (openintro) | `heart_transplant` (openintro) | `flights` (nycflights13)

StatLens Exercises

The following exercises use StatLens, an interactive statistics tool. Each exercise includes a link that opens a pre-loaded dataset — explore, interact, and answer the questions.

1. **Which chart for categorical data?** Open the jury racial composition data in StatLens. This dataset records the race of 275 jurors from a county. The chart opens with numeric labels hidden so you can focus on the visual patterns.

   a. Start with the **bar chart**. Using only the bar heights, rank the four racial groups from most to least represented. Which group is clearly the largest?

   b. Switch to a **pie chart**. Try to rank the four groups from largest to smallest using only the visual size of each slice. Is it easier or harder than with the bar chart? Check the **Show values** box to reveal the numbers — how close was your ranking?

   c. Switch to a **waffle chart** (uncheck **Show values** first). Count the squares for the White group to estimate their percentage. Then check **Show values** to see the actual number. Is estimating percentages easier from the waffle chart or the pie chart? Why?

   d. If a community advocacy group wanted to argue that the jury pool does not reflect the racial makeup of the county, which chart type would make the strongest visual case? Explain your reasoning.

2. **Detecting association in a two-way table.** Open the heart transplant data in StatLens. This dataset records whether patients in a heart transplant study were in the treatment group (received a transplant) or the control group, and whether they survived or died.

   a. Look at the **contingency table** (counts). How many patients received a transplant? How many survived overall? How many patients both received a transplant and survived?

   b. Switch the table to **Row Proportions**. What proportion of the treatment group survived? What proportion of the control group survived? Based on these proportions, does there appear to be an association between receiving a transplant and survival?

   c. Now switch to **Column Proportions**. Among those who survived, what proportion were in the treatment group? Why does this answer a different question than what you found in part (b)?

   d. View the data as a **100% Stacked** bar chart. If group assignment and survival were independent (not associated), what would the bar chart look like? Describe how the actual chart differs from that independent scenario.

Chapter 4

Exploring Quantitative Data

In the previous chapter we explored categorical data. Now we turn to **quantitative data** — variables that take numerical values where arithmetic makes sense (heights, incomes, interest rates). This chapter teaches you to both *see* and *measure* distributions. We first **visualize** a distribution with dot plots and histograms, then **measure** its center and spread, flag any **outliers**, and pull it all together with a **box plot**. By the end, you will be able to describe any distribution in terms of its shape, center, spread, and unusual features.

4.1 Visualizing quantitative data

4.1.1 Dot plots

A **dot plot** is the simplest display for a single quantitative variable. Each observation is represented by a dot placed above its value on a number line. When observations share the same value (or are rounded to the same value), the dots stack vertically.

Consider the interest rates on 50 loans from a lending platform. In a dot plot, each of the 50 loans is a single dot. The red triangle marks the mean.

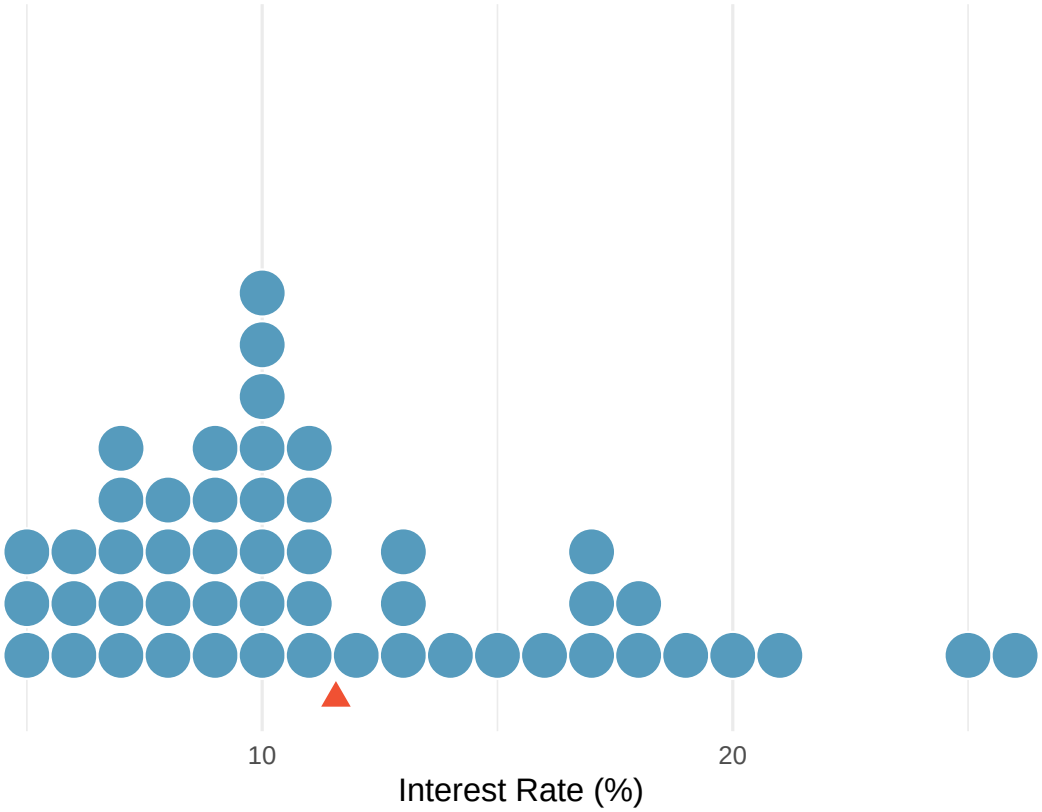


Figure 4.1: A dot plot of interest rates for 50 loans. Most values cluster between 5% and 15%, with a few high rates pulling the distribution to the right. The red triangle marks the mean (11.57%). Open in StatLens

Dot plots are most useful for small datasets (up to about 50–100 observations) where you want to see every individual value. For larger datasets, histograms provide a better summary.

4.1.2 Histograms

A **histogram** groups quantitative data into bins (intervals) and displays the count or proportion of observations in each bin as a bar. Unlike bar charts for categorical data, histogram bars touch each other because the bins represent a continuous range of values.

To build a histogram, we:

1. Choose a bin width (e.g., 2.5 percentage points for interest rates)
2. Count how many observations fall in each bin
3. Draw bars whose heights equal the counts

Interest rate bin	Count
(5.0% - 7.5%]	11
(7.5% - 10.0%]	15
(10.0% - 12.5%]	8
(12.5% - 15.0%]	4
(15.0% - 17.5%]	5
(17.5% - 20.0%]	4
(20.0% - 22.5%]	1

Interest rate bin	Count
(22.5% - 25.0%]	1
(25.0% - 27.5%]	1

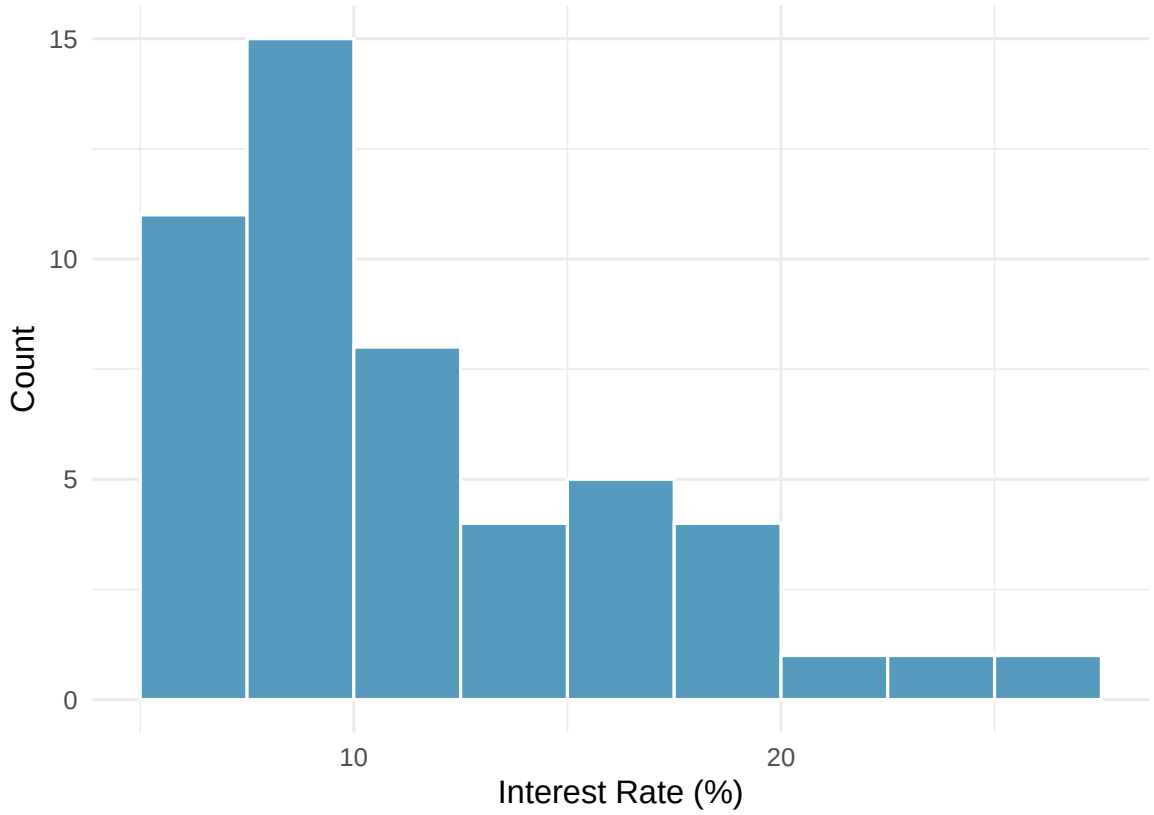


Figure 4.2: A histogram of interest rates with bin width 2.5 percentage points. The distribution is right skewed. Open in StatLens

Histograms provide a view of the **data density**: higher bars indicate where data values are more common, and lower bars indicate where they are less common.

Choosing bin width. The bin width affects the appearance of a histogram. Too few bins (wide bins) can obscure important features; too many bins (narrow bins) can create a jagged display that is hard to interpret. There is no single "correct" bin width — experiment with different choices and look for the display that best reveals the shape of the data.

4.1.3 Density plots

A **density plot** is a smoothed version of a histogram. Instead of bars, it draws a smooth curve that estimates the distribution of the data. The area under the entire curve equals 1.

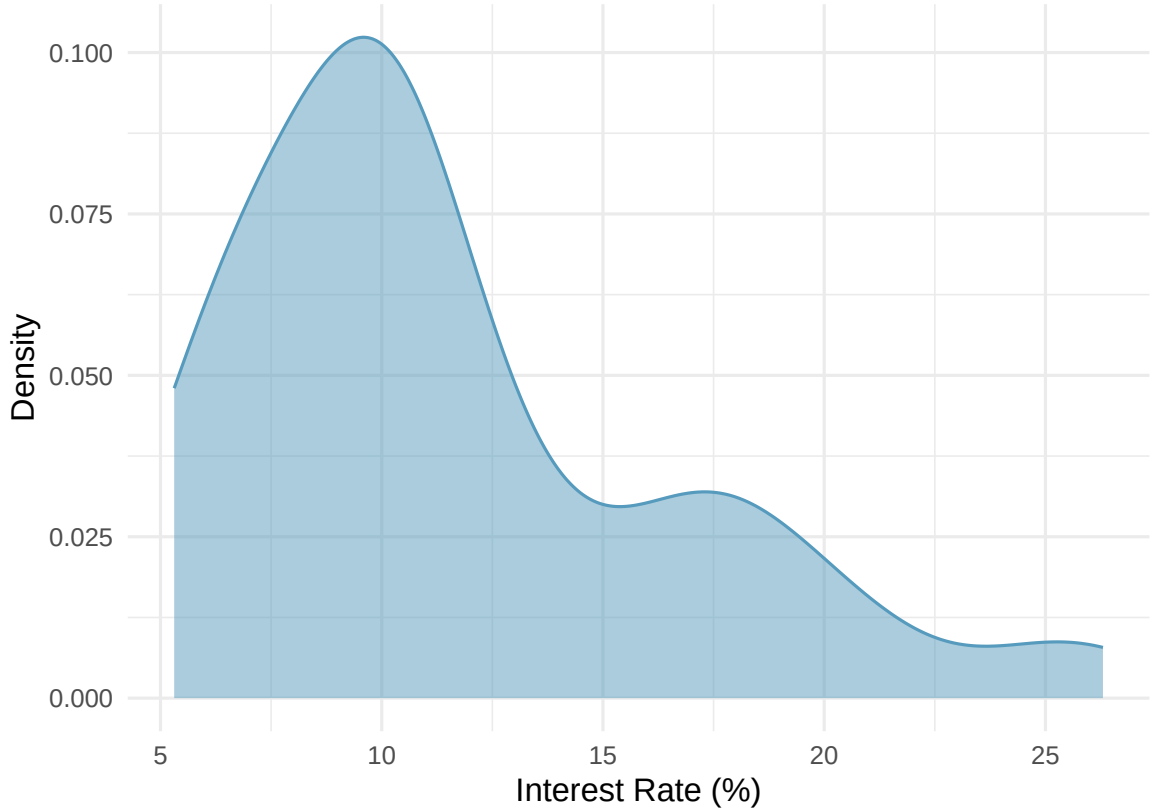


Figure 4.3: A density plot of interest rates, showing the same right-skewed shape as the histogram. Open in StatLens

Density plots are particularly useful for comparing the shapes of distributions because overlapping curves are easier to read than overlapping histogram bars. Figure 4.4 overlays the birth-weight distributions for babies of nonsmoking and smoking mothers. Both curves have a similar bell-like shape, but the smokers' curve sits slightly to the left — a lower average birth weight — a shift that is much easier to see with two overlaid curves than it would be with two sets of competing histogram bars.

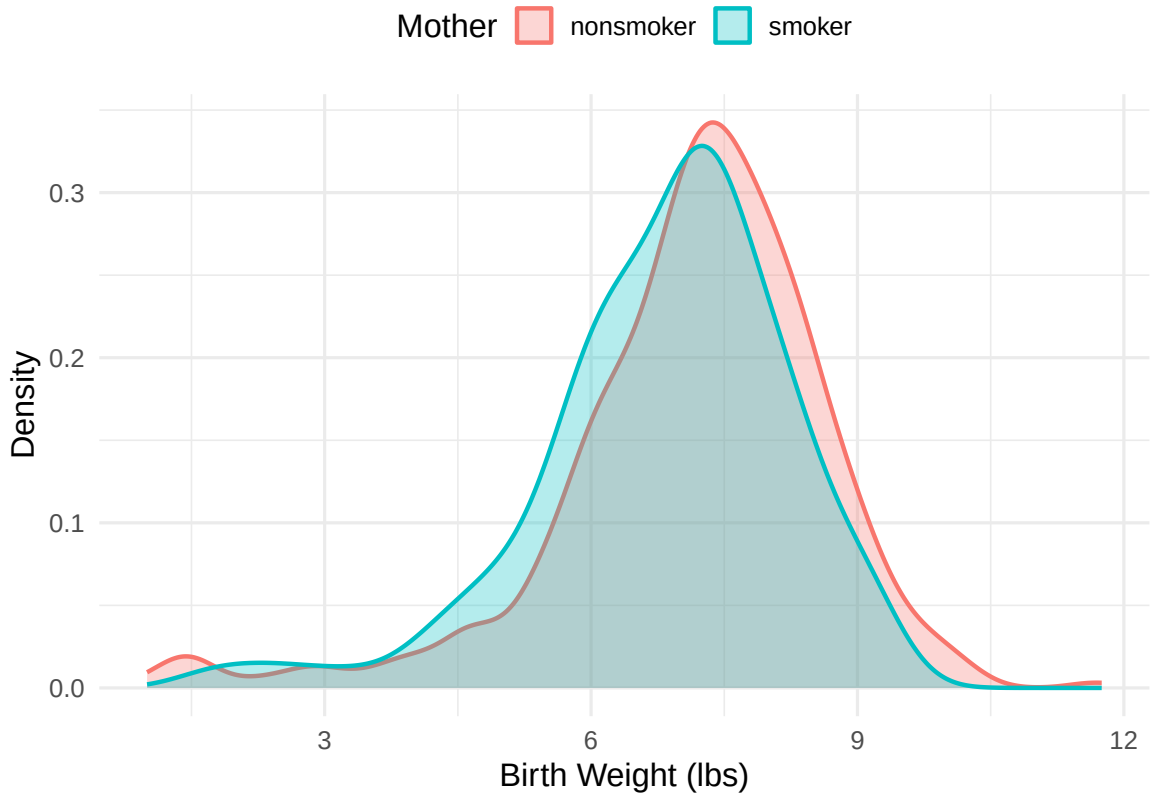


Figure 4.4: Birth-weight density curves for babies of nonsmoking and smoking mothers, overlaid on a shared axis. The smokers' curve is shifted slightly toward lower weights. Open in StatLens

4.1.4 Skewness

Histograms are especially useful for identifying the **shape** of a distribution. One key characteristic is symmetry.

A distribution is **symmetric** if the left and right sides are approximately mirror images of each other.

A distribution is **right skewed** (or skewed to the right) if the right **tail** — the thinner end of the distribution — is longer. Most observations are on the left, with a few large values pulling the tail to the right.

A distribution is **left skewed** (or skewed to the left) if the left tail is longer. Most observations are on the right, with a few small values pulling the tail to the left.

Identifying skew. The direction of skew refers to the direction of the *long tail*, not the direction where most data are concentrated. A distribution with a long tail stretching to the right is right skewed, even though the bulk of the data is on the left side.

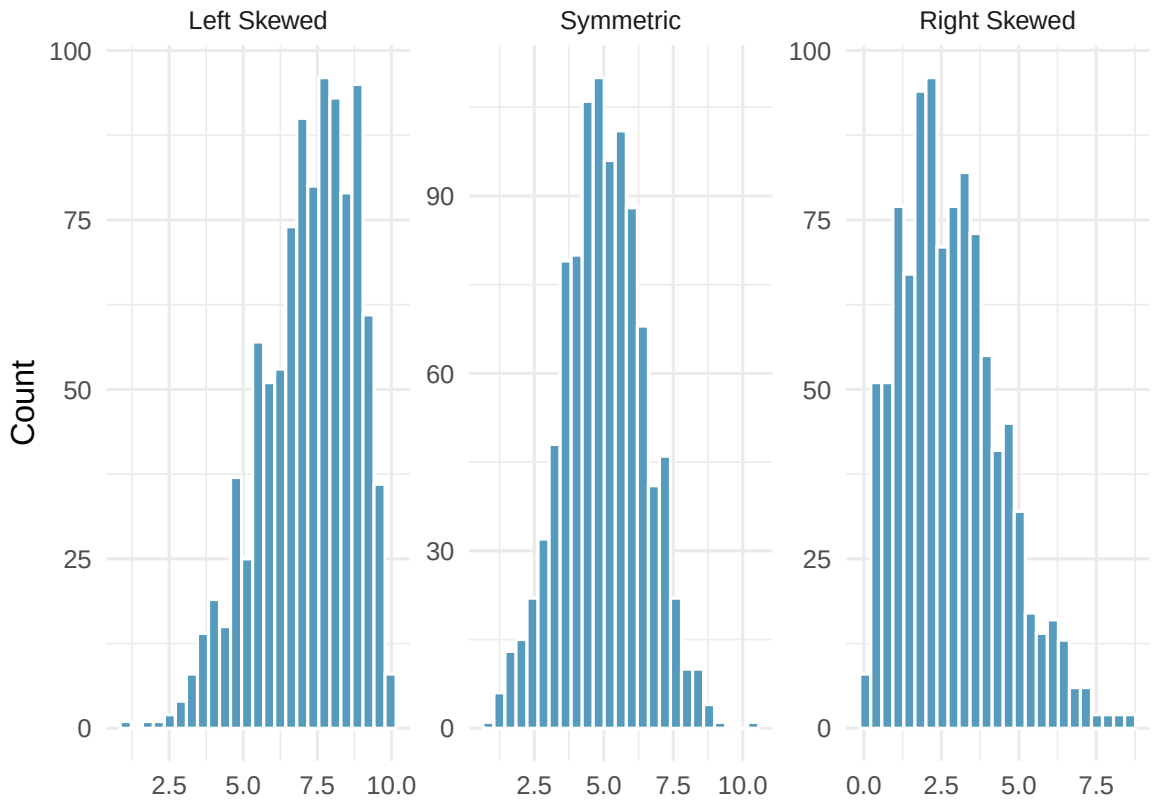


Figure 4.5: Three distributions illustrating left skewed, symmetric, and right skewed shapes.

Consider the distribution of home prices in a city. Would you expect this distribution to be symmetric, left skewed, or right skewed? Explain your reasoning.

Show answer

Home prices tend to be right skewed. Most homes fall in a moderate price range, but a small number of very expensive homes create a long right tail. This is common with income, wealth, and price data.

4.1.5 Modality

A **mode** is a prominent peak in a distribution. A distribution with a single prominent peak is called **unimodal**. A distribution with two prominent peaks is **bimodal**, and a distribution with more than two prominent peaks is **multimodal**.

A distribution with no prominent peaks — where all values are roughly equally likely — is called **uniform**.

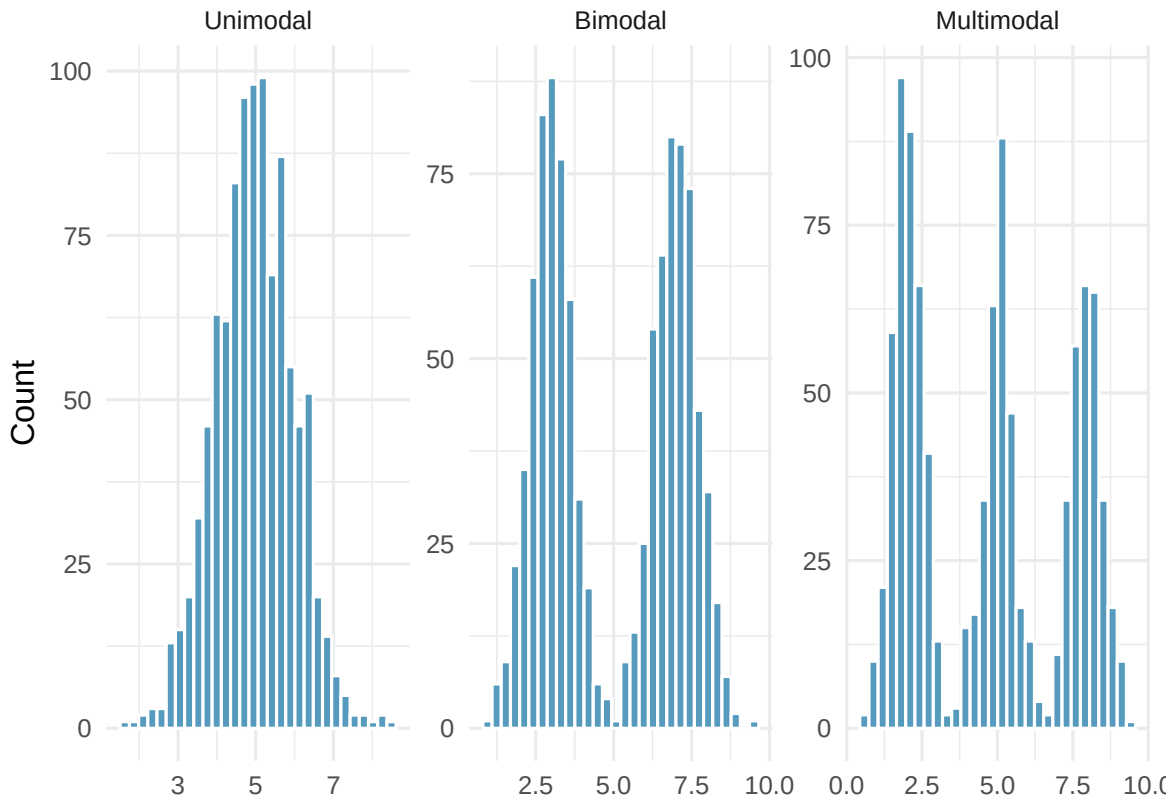


Figure 4.6: Three distributions: unimodal, bimodal, and multimodal.

Height measurements were collected from a group of young students and adult teachers at an elementary school. How many modes would you expect in a histogram of these heights?

You would likely expect two modes (bimodal): one centered on the typical height of children and another centered on the typical height of adults.

A histogram of interest rates shows one prominent peak near 7%. There is a very small bump near 20%, but it differs from its neighboring bins by only one or two observations. Would you call this distribution unimodal or bimodal?

Show answer

Unimodal. The small bump near 20% is not a prominent peak — it differs only slightly from its neighbors. We look for prominent peaks when identifying modes.

4.1.6 Transforming skewed data (optional)

This topic is optional; later chapters do not depend on it.

Strongly right-skewed variables — incomes, populations, prices — can be hard to read on their original scale, because a long tail of large values squeezes most of the data into a narrow band on the left. A **transformation** re-expresses the data on a new scale to make the distribution easier to see and describe. The most common choice is the **logarithm**.

Taking the log of each value pulls in the long right tail and often makes a right-skewed distribution roughly symmetric. Figure 4.7 shows the annual incomes of the 50 loan applicants on the original dollar scale and after a base-10 log transformation.

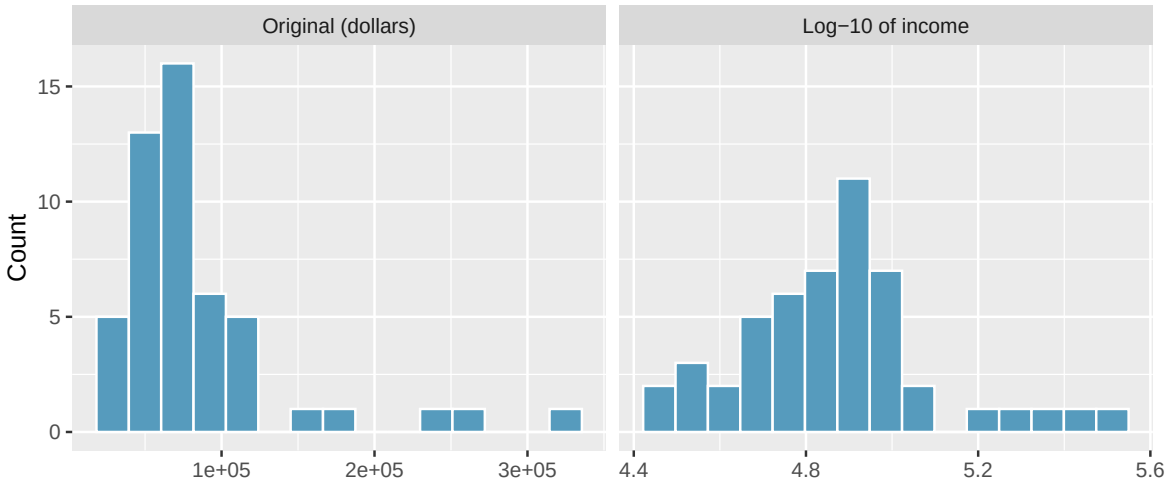


Figure 4.7: Annual income of 50 loan applicants, shown on the original dollar scale (left) and after a base-10 logarithm (right). The log transformation pulls in the long right tail, leaving a more symmetric distribution. Open in StatLens

On the log scale a one-unit increase corresponds to a *tenfold* increase in income (because we used base 10), so the transformed values are easier to compare and summarize. After transforming, you can describe shape, center, and spread on the new scale just as you would for any distribution — but remember that those summaries now describe the *log* of income, not income in dollars.

When (and when not) to transform. A log transformation is useful for strongly right-skewed, positive-valued data such as incomes, city populations, or reaction times. It cannot be applied to values that are zero or negative, and it is rarely needed for roughly symmetric data. A square-root transformation is a milder alternative for moderate skew. Whether to transform depends on your goal: transformations help reveal structure, but the original scale is usually what readers ultimately care about.

4.2 Measures of center

4.2.1 The mean

The **mean** (also called the **average**) is computed by adding up all the observed values and dividing by the number of observations.

Mean.

The sample mean can be calculated as the sum of the observed values divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

The notation $\bar{x}$ (read "x-bar") represents the sample mean. It's useful to think of the mean as the **balancing point** of a distribution. In the dot plot above, the mean of 11.57% is where the data would

balance if placed on a seesaw — the few high interest rates on the right pull the balance point away from where most of the dots are concentrated.

The population mean is denoted by the Greek letter μ (pronounced "mew"). We often cannot measure μ directly because we rarely have access to every member of the population. Instead, we estimate μ using the sample mean $\bar{x}$.

In the formula for the mean, what does x_1 correspond to? And x_2? What does x_i represent in general?

Show answer

x_1 is the value of the variable for the first observation in the dataset, x_2 is the value for the second observation, and x_i is the value for the ith observation. For example, if the variable is interest rate, x_4 is the interest rate of the fourth loan.

Suppose we want to compute the average income per person in the US. We might think to take the mean of per capita incomes across all 3,142 counties. What is wrong with this approach?

Each county represents a different number of people. If we simply average the county-level incomes, we treat a county with 5,000 residents the same as one with 5,000,000 residents. Instead, we should compute the total income for each county, add up all counties' totals, and divide by the total population. This is called a **weighted mean**. Failing to weight by population would underestimate the true average because larger, higher-income urban counties would be underrepresented.

4.2.2 The median

The **median** is the value in the middle of the ordered data. It divides the dataset in half: 50% of the observations fall below the median and 50% fall above it.

Median: the number in the middle.

If the data are ordered from smallest to largest, the **median** is the observation right in the middle. If there are an even number of observations, the median is the average of the two middle values.

The dot plot below shows both the mean and the median marked on the interest rate data. Notice that the median (9.93%) is lower than the mean (11.57%) — the few loans with very high interest rates pull the mean to the right, but the median stays anchored in the middle of the data.

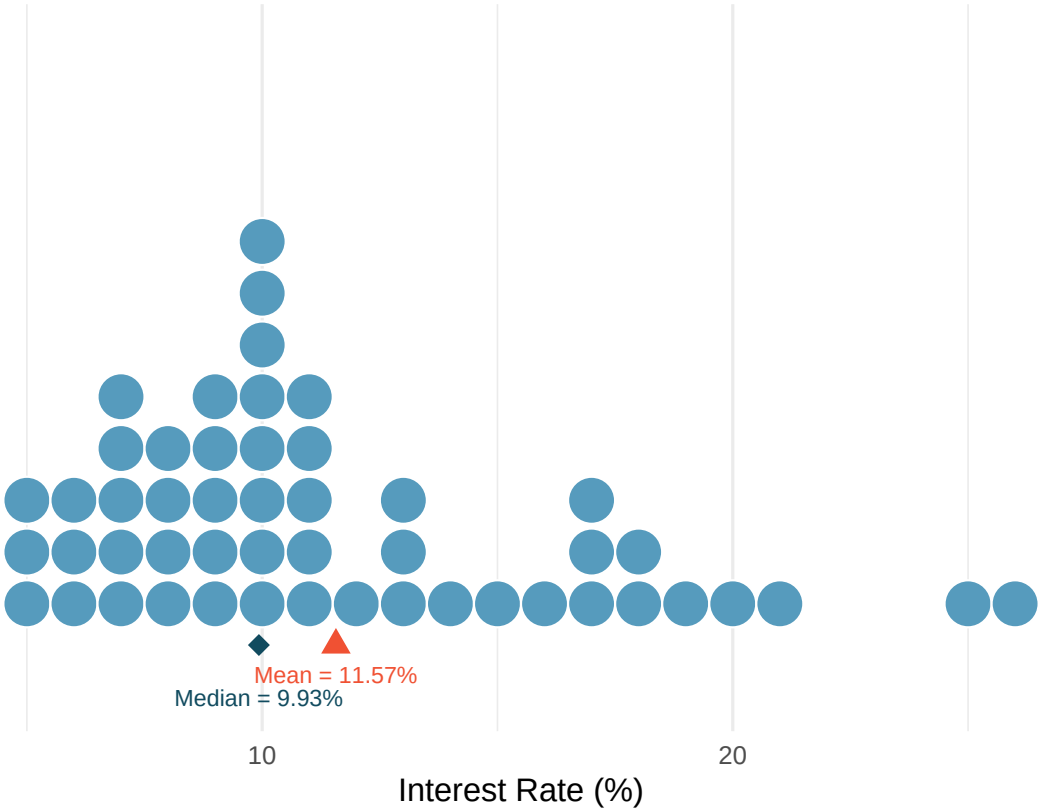


Figure 4.8: A dot plot of interest rates with the mean (red triangle, 11.57%) and median (blue diamond, 9.93%) marked. The mean is pulled to the right by the high-interest loans. Open in StatLens

For a dataset with 7 observations arranged in order, which observation is the median?

Show answer

The 4th observation. With an odd number of observations, the median is the single middle value. There are 3 observations below it and 3 above it.

4.2.3 The mode

The **mode** is the most frequently occurring value in a dataset. While useful for categorical data (e.g., the most common eye color), the mode is less useful for continuous quantitative data because many values may occur only once.

In practice, when we refer to the "mode" of a quantitative distribution, we usually mean the location of a prominent *peak* in a histogram or density plot (as discussed in Section 4.1), rather than a single repeated value.

4.2.4 How skewness affects the mean and median

Now we can connect what we see in a histogram to the sample statistics we computed earlier.

In a **symmetric** distribution, the mean and median are approximately equal.

In a **right-skewed** distribution, the mean is pulled toward the right tail and is larger than the median.

In a **left-skewed** distribution, the mean is pulled toward the left tail and is smaller than the median.

Relationship between mean, median, and skewness.

- Right skewed: mean $>$ median
- Symmetric: mean $\approx$ median
- Left skewed: mean $<$ median

The direction of skew pulls the mean toward the tail. The median is more resistant to this pull.

The interest rate distribution is right skewed, with a mean of 11.57% and a median of 9.93%. How does the skewness explain the gap between these two values?

The few loans with very high interest rates (above 20%) pull the mean to the right. The median, which simply finds the middle value, is not affected by how extreme those high rates are. This is why the mean (11.57%) is larger than the median (9.93%) in a right-skewed distribution.

See it in action. Open the Descriptive Statistics Explorer on the interest-rate data — switch between histograms, dot plots, and box plots, and compare the mean and median in the summary panel.

4.2.5 When to use the mean vs. the median

The mean and median measure different things. The choice between them depends on the context:

A company wants to understand its profit per customer. Should it use the mean or the median?

It depends on the question:

- If the concern is overall profitability (total profit divided by number of customers), the **mean** is the right statistic. A company could have a positive median profit per customer and still be unprofitable if a few customers generate large losses.
- If the concern is understanding what the *typical* customer experience looks like, the **median** tells you more about the profit from a representative customer.

Mean vs. median: which to choose?

- The **mean** is best when you care about the *total* or *aggregate* (e.g., total revenue, total time spent).
- The **median** is best when you care about the *typical* observation, especially when the distribution is skewed or contains outliers.
- For symmetric distributions without extreme outliers, the mean and median are approximately equal and either is appropriate.

The distribution of home prices in a city is strongly right skewed. If you wanted to describe the price of a typical home, should you report the mean or the median? Why?

Show answer

The median. In a right-skewed distribution, a few very expensive homes pull the mean upward, making it unrepresentative of a typical home price. The median is not affected by these extreme values and better represents what a typical buyer would encounter.

4.3 Measures of spread

Knowing the center of a distribution is important, but it tells only part of the story. Two distributions can have the same center but look very different if one is tightly clustered and the other is widely dispersed. We need measures of **variability** (also called spread or dispersion).

4.3.1 Range

The simplest measure of spread is the **range**: the difference between the maximum and minimum values.

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

While easy to compute, the range has a major weakness: it depends entirely on the two most extreme observations. A single outlier can dramatically increase the range, giving a misleading impression of the overall variability. For this reason, we usually prefer the IQR or standard deviation.

4.3.2 The interquartile range (IQR)

The **interquartile range** measures the spread of the middle 50% of the data. It will reappear as the length of the box when we construct box plots later in this chapter.

Interquartile range (IQR).

The IQR is computed as:

$$\text{IQR} = Q_3 - Q_1$$

where Q_1 is the first quartile (25th percentile) and Q_3 is the third quartile (75th percentile). To find them by hand, split the ordered data at the median: Q_1 is the median of the lower half and Q_3 is the median of the upper half. When the number of observations is odd, the median itself is left out of both halves.

For the 50-loan interest rate data, $Q_1 = 7.96\%$ and $Q_3 = 13.72\%$. What is the IQR, and what does it tell us?

IQR $= 13.72\% - 7.96\% = 5.76$ percentage points. This means the middle 50% of loans have interest rates spanning a range of about 6 percentage points. Half of all loans have rates between about 8% and 14%.

4.3.3 Variance and standard deviation

The **standard deviation** is the most commonly used measure of spread. It roughly describes how far the typical observation is from the mean.

We start with the concept of a **deviation**: the distance of an observation from the mean.

$$\text{deviation of } x_i = x_i - \bar{x}$$

Some deviations are positive (observations above the mean) and some are negative (observations below the mean). If we simply averaged the deviations, the positives and negatives would cancel out, giving us zero. To avoid this, we *square* each deviation before averaging.

The **variance** (denoted s^2) is the average of the squared deviations, using $n - 1$ in the denominator:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

The **standard deviation** (denoted s) is the square root of the variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

We divide by $n-1$ rather than n when computing a sample's variance. The mathematical reasons are beyond this course, but the practical effect is that this makes the statistic slightly more reliable as an estimate of the population variance.

Why do we square the deviations instead of just taking the absolute value? What are the two consequences of squaring?

Show answer

Squaring the deviations does two things: (1) it eliminates negative signs (since any number squared is non-negative), and (2) it makes large deviations relatively *much* larger than small ones. An observation that is 10 units from the mean contributes 100 to the sum of squared deviations, while one that is 2 units away contributes only 4. This means the variance and standard deviation are especially sensitive to outliers.

The following histogram shows the interest rate data with shaded bands marking one and two standard deviations from the mean. The standard deviation of 5.05 percentage points gives us a sense of how spread out the data are around the mean of 11.57%.

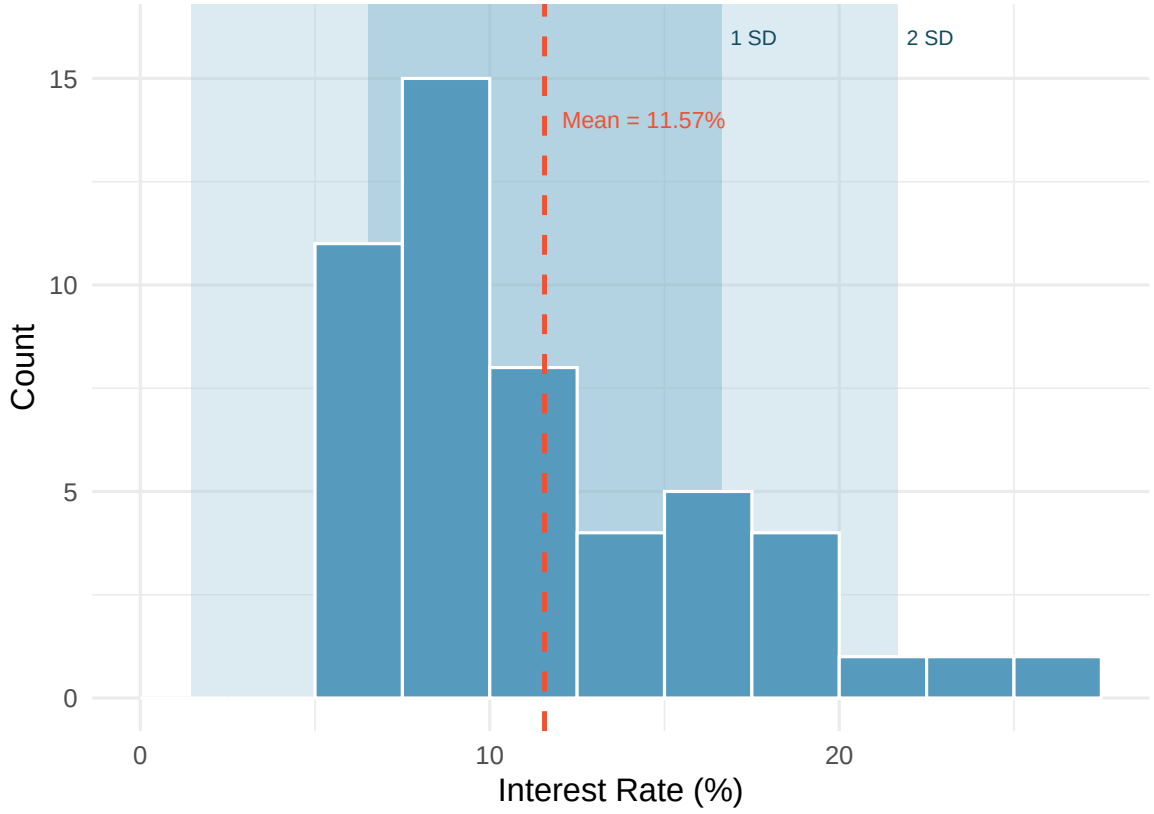


Figure 4.9: A histogram of interest rates with shaded regions showing one standard deviation (dark) and two standard deviations (light) from the mean. Open in StatLens

For the 50-loan interest rate data, the variance is $s^2 = 25.52$ and the standard deviation is $s =$

$\sqrt{25.52} = 5.05$ percentage points. What does a standard deviation of 5.05 percentage points mean in context?

The standard deviation tells us that interest rates typically deviate about 5 percentage points from the mean (11.57%). Most loans have rates that are within one standard deviation of the mean — roughly between 6.5% and 16.6%.

Like the mean, the population values for variance and standard deviation have special symbols: σ^2 for the variance and σ for the standard deviation. The Greek letter σ is pronounced "sigma."

Variance and standard deviation.

The variance is the average squared distance from the mean. The standard deviation is the square root of the variance. The standard deviation represents the *typical* deviation of observations from the mean. Often about 68% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, these percentages are not strict rules.

4.4 Outliers and robust statistics

4.4.1 Identifying outliers

An **outlier** is an observation that appears extreme relative to the rest of the data.

Examining data for outliers serves several purposes:

- Identifying strong skew in the distribution
- Identifying possible data collection or data entry errors
- Providing insight into interesting or unusual properties of the data

Not all outliers are errors, and not all errors are outliers. Some datasets naturally have long tails and outlying points are legitimate observations. **Never remove an outlier without a substantive reason** (e.g., confirmed data entry error). Blindly removing outliers can lead to misleading conclusions.

A common rule for identifying outliers in the context of box plots is the **1.5 IQR rule**: any observation more than $1.5 \times$ IQR below Q_1 or above Q_3 is flagged as a potential outlier. This is the rule used to determine which points appear as dots beyond the whiskers of a box plot.

4.4.2 Robust statistics

A statistic is called **robust** (or **resistant**) if extreme observations have little effect on its value.

Consider what happens when we change a single extreme observation in the interest rate data. The table below shows how the median, IQR, mean, and standard deviation respond when the most extreme value (26.3%) is moved:

Scenario	Median	IQR	Mean	SD
Original data	9.93	5.76	11.57	5.05
Move 26.3% to 15%	9.93	5.76	11.34	4.61
Move 26.3% to 35%	9.93	5.76	11.74	5.68

In the table above, the median and IQR are unchanged across all three scenarios, but the mean and standard deviation shift. Why?

The median and IQR are determined by the values near the middle of the ordered data (Q_1, the median, and Q_3). Moving a single extreme value does not change the positions of these middle values. The mean and standard deviation, however, use every data value in their calculations. The mean adds up all values, so changing one extreme value changes the sum. The standard deviation squares deviations from the mean, so extreme values have an outsized effect.

Robust vs. non-robust statistics.

Robust (resistant to outliers)	Not robust (sensitive to outliers)
Median	Mean
IQR	Standard deviation
	Range

When a distribution is skewed or contains outliers, robust statistics (median and IQR) often give a more representative summary of the "typical" observation. When a distribution is roughly symmetric with no extreme outliers, the mean and standard deviation work well and carry additional mathematical advantages.

The distribution of salaries at a company is right skewed because a few executives earn much more than most employees. Which pair of sample statistics would better represent the "typical" salary and its variability: (mean, standard deviation) or (median, IQR)?

Show answer

The median and IQR. In a right-skewed distribution, the mean is pulled upward by the high executive salaries, making it unrepresentative of the typical employee. The median is not affected by these extreme values, and the IQR captures the spread of the middle 50% of salaries.

The distribution of loan amounts is right skewed. If you wanted to understand the typical loan size, should you be more interested in the mean or the median?

Show answer

If the goal is to understand what a typical individual loan looks like, the median is probably more useful. However, if the goal is to understand something that scales — such as the total amount of money needed on hand to fund 1,000 loans — then the mean would be more useful.

See it in action. Open the Outliers and Robust Statistics activity on US county populations — one of the most skewed distributions you'll encounter — to see firsthand why the mean and SD mislead and the median and IQR don't.

4.5 Box plots

A **box plot** (also called a box-and-whisker plot) provides a compact summary of a distribution using five numbers: the minimum, first quartile (Q_1), median, third quartile (Q_3), and maximum — with special handling for outliers.

4.5.1 Constructing a box plot

The components of a box plot are:

1. **The box**: Spans from Q_1 (the 25th percentile) to Q_3 (the 75th percentile). This box contains the middle 50% of the data. The length of the box is the **interquartile range** (IQR = $Q_3 - Q_1$).

2. **The median line**: A line inside the box marks the median (the 50th percentile), which divides the data in half.
3. **The whiskers**: Lines extending from the box toward the smallest and largest observations that are *not* outliers. Specifically, the whiskers extend to the most extreme data points within $1.5 \times$ IQR of the box.
4. **Outlier points**: Any observations beyond $1.5 \times$ IQR from the edges of the box are plotted individually as dots. These are potential outliers.

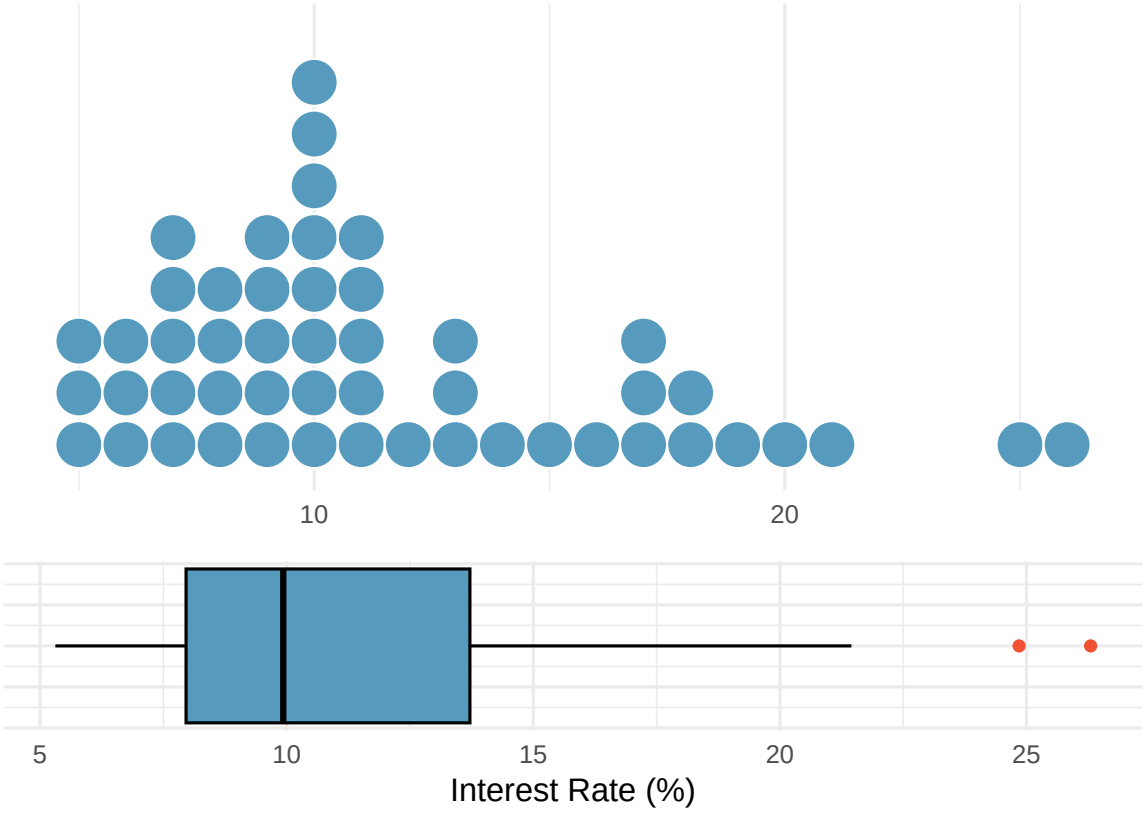


Figure 4.10: A dot plot and box plot of the same interest rate data, shown one above the other for comparison. The box plot summarizes the distribution with five numbers plus outliers. Open in StatLens

The **first quartile** (Q_1) is the 25th percentile — 25% of the data fall below this value. The **third quartile** (Q_3) is the 75th percentile — 75% of the data fall below this value.

The **interquartile range** (IQR) is the range of the middle 50% of the data: IQR $= Q_3 - Q_1$.

4.5.2 Interpreting a box plot

A box plot of interest rates shows: the left edge of the box at about 8%, the median line at about 10%, and the right edge of the box at about 14%. Two dots appear beyond the right whisker at about 25% and 26%. What can we learn from this display?

The middle 50% of interest rates fall between about 8% and 14% (the IQR is about 6 percentage points). The median rate is about 10%. The right whisker extends to about 21%, and the two dots beyond the

whisker at 25% and 26% indicate unusually high interest rates. The box plot also reveals right skew, since the median is closer to the left edge of the box and the right whisker is longer than the left whisker.

What percentage of observations fall between Q_1 and the median? What percentage fall between the median and Q_3?

Show answer

25% of the data fall between Q_1 and the median, and another 25% fall between the median and Q_3. Together, the middle 50% of the data lie within the box.

See it in action. Open the Comparing Charts for Quantitative Data activity to walk through dot plots, histograms, and box plots on home sale prices and see what each chart type reveals (and hides).

4.6 Describing a distribution

A complete description of a distribution should address three things: **shape**, **center**, and **spread**.

A good description of a distribution should always:

1. Name the **shape**: symmetric or skewed (and direction), and the number of modes
2. Give the **center**: report the mean and/or median with units
3. Give the **spread**: report the standard deviation and/or IQR with units
4. Note any **unusual features**: outliers, gaps, clusters
5. Use **context**: refer to the actual variable and its units, not just abstract numbers

Example: "The distribution of commute times is unimodal and right skewed, with a median of 22 minutes and an IQR of 15 minutes. A few commuters have commute times exceeding 90 minutes."

Write a complete description of the interest rate distribution from the loan dataset.

The distribution of interest rates is **unimodal** and **right skewed**. The **center** is around 10–12% (median = 9.93%, mean = 11.57%). The **spread** is moderate: the standard deviation is 5.05 percentage points and the IQR is about 6 percentage points. There are a few **unusually high** interest rates above 20%, which contribute to the right skew and the gap between the mean and median.

A description says: "The mean is 15 and the standard deviation is 3." What is missing from this description?

Show answer

The description lacks context (what variable? what units?), shape information (symmetric? skewed? unimodal?), and any mention of unusual features. A better version might be: "The distribution of wait times (in minutes) is approximately symmetric and unimodal, with a mean of 15 minutes and a standard deviation of 3 minutes. There are no obvious outliers."

4.7 The empirical rule and Chebyshev's theorem

4.7.1 The empirical rule (68-95-99.7 rule)

For distributions that are approximately **bell-shaped** (symmetric and unimodal), the standard deviation provides a useful ruler. The most important bell-shaped model is the **normal distribution** — it comes up so often in statistics that it has its own name. We'll study it formally later. Not every

bell-shaped distribution is exactly normal, but the rule below works well for any roughly bell-shaped distribution.

The empirical rule.

For a roughly bell-shaped distribution:

- About **68%** of the data fall within 1 standard deviation of the mean: $(\bar{x} - s,\ \bar{x} + s)$
- About **95%** of the data fall within 2 standard deviations of the mean: $(\bar{x} - 2s,\ \bar{x} + 2s)$
- About **99.7%** of the data fall within 3 standard deviations of the mean: $(\bar{x} - 3s,\ \bar{x} + 3s)$

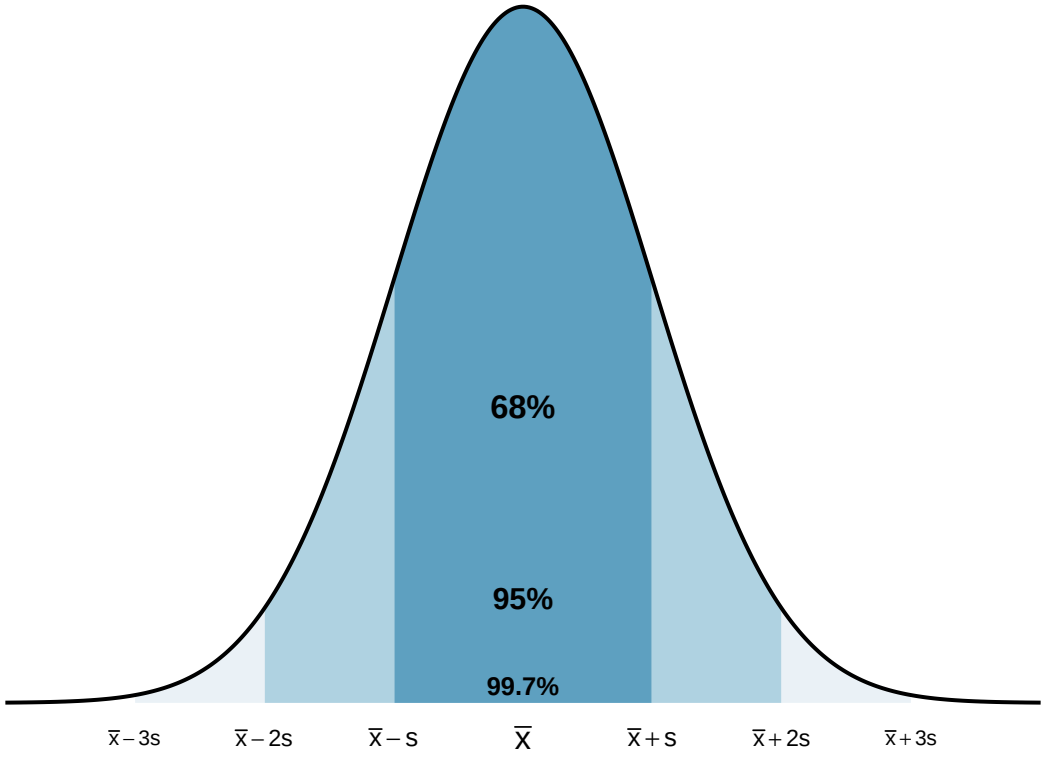


Figure 4.11: A bell-shaped distribution with shaded regions showing 68%, 95%, and 99.7% of the data within 1, 2, and 3 standard deviations of the mean, respectively. ↗ Try this live — drag the cutlines in the Normal Distribution tool to see the 68-95-99.7 boundaries move as you change μ and σ.

The heights of adult women in the US are approximately bell-shaped with a mean of 64.5 inches and a standard deviation of 2.5 inches. Use the empirical rule to estimate the range of heights for the middle 95% of women.

By the empirical rule, about 95% of the data fall within 2 standard deviations of the mean:

$$64.5 - 2(2.5) = 59.5 \text{ inches} \quad \text{to} \quad 64.5 + 2(2.5) = 69.5 \text{ inches}$$

So about 95% of adult women in the US are between 59.5 inches (4'11.5") and 69.5 inches (5'9.5") tall.

Using the same height distribution (mean = 64.5 inches, SD = 2.5 inches), what percentage of women are shorter than 59.5 inches?

Show answer

By the empirical rule, about 95% of women are between 59.5 and 69.5 inches. That means about 5% are outside this range. Since the distribution is symmetric, about half of that 5% (i.e., about 2.5%) are below 59.5 inches.

The empirical rule is only an approximation, and it works best for distributions that are roughly bell-shaped. For skewed or multimodal distributions, the percentages can be quite different. For example, in a right-skewed distribution, more than 68% of the data may fall within one standard deviation of the mean because the bulk of the data is concentrated on one side.

Three distributions all have the same mean (0) and the same standard deviation (1), but one is bimodal, one is bell-shaped, and one is right skewed. Would the empirical rule work well for all three? Why or why not?

Show answer

No. The empirical rule works well only for the bell-shaped distribution. The bimodal distribution would have more data near the edges (near the two peaks) and less in the middle. The right-skewed distribution would have more data close to the mean on one side and a long tail on the other. In both cases, the 68-95-99.7 percentages would be poor approximations.

4.7.2 Chebyshev's theorem

While the empirical rule applies only to bell-shaped distributions, **Chebyshev's theorem** provides a guarantee that works for *any* distribution, regardless of shape:

Chebyshev's theorem.

For **any** distribution (regardless of shape), at least $\left(1 - \frac{1}{k^2}\right) \times 100\%$ of the data fall within k standard deviations of the mean, for any $k > 1$.

Standard deviations (k)	At least this % of data
2	75%
3	89%
4	93.75%

Chebyshev's theorem is weaker than the empirical rule (it guarantees "at least 75%" within 2 standard deviations, while the empirical rule says "about 95%" for bell-shaped data), but it has the advantage of applying to every distribution. It is most useful when you cannot assume the distribution is bell-shaped.

4.8 Chapter review

4.8.1 Summary

This chapter introduced the tools for exploring quantitative data — both visual and numerical. **Dot plots** show individual values for small datasets, **histograms** reveal the shape and density of a distribution, **density plots** provide smoothed alternatives, and **box plots** give a compact five-number summary. Distribution shape is described using **symmetry** (left skewed, symmetric, right skewed) and **modality** (unimodal, bimodal, multimodal). The **mean** and **median** measure center, with the

mean being the balancing point and the median being the middle value. The **standard deviation** and **IQR** measure spread, with the standard deviation roughly representing the typical deviation from the mean and the IQR representing the range of the middle 50%. The **median** and **IQR** are robust to outliers, while the **mean** and **standard deviation** are sensitive to extreme values. The **empirical rule** connects the standard deviation to the distribution shape for bell-shaped data: about 68%, 95%, and 99.7% of observations fall within 1, 2, and 3 standard deviations of the mean.

4.8.2 Key terms

- Dot plot, histogram, density plot, box plot
- Data density, bin width
- Symmetric, left skewed, right skewed, tail
- Unimodal, bimodal, multimodal, uniform
- Mean ($\bar{x}$), median, mode
- Population mean (μ), weighted mean
- Deviation, variance (s^2), standard deviation (s)
- Population variance (σ^2), population standard deviation (σ)
- Range, interquartile range (IQR)
- First quartile (Q_1), third quartile (Q_3), percentile
- Outlier, robust statistic (resistant statistic)
- Empirical rule (68-95-99.7 rule), Chebyshev's theorem

4.9 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Make-up exam.** In a class of 25 students, 24 of them took an exam in class and 1 student took a make-up exam the following day. The professor graded the first batch of 24 exams and found an average score of 74 points with a standard deviation of 8.9 points. The student who took the make-up the following day scored 64 points on the exam.

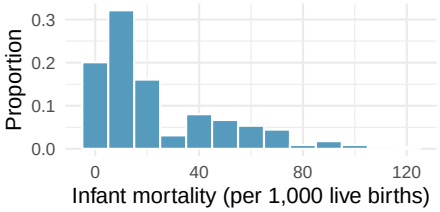
a. Does the new student's score increase or decrease the average score?

b. What is the new average?

c. Does the new student's score increase or decrease the standard deviation of the scores?

2. **Infant mortality.** The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates for 224 countries for which such data were available in 2014.

   a. Estimate Q1, the median, and Q3 from the histogram.
   b. Would you expect the mean of this dataset to be smaller or larger than the median? Explain your reasoning.



3. **Days off at a mining plant.** Workers at a particular mining site receive an average of 35 days paid vacation, which is lower than the national average. The manager of this plant is under pressure from a local union to increase the amount of paid time off. However, he does not want to give more days off to the workers because that would be costly. Instead he decides he should fire 10 employees in such a way as to raise the average number of days off that are reported by his employees. In order to achieve this goal, should he fire employees who have the most number of days off, least number of days off, or those who have about the average number of days off?

4. **Medians and IQRs.** For each part, compare distributions A and B based on their medians and IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make sure to explain your reasoning. *Hint:* It may be useful to sketch dot plots of the distributions.

a. **A:** 3, 5, 6, 7, 9; **B:** 3, 5, 6, 7, 20

b. **A:** 3, 5, 6, 7, 9; **B:** 3, 5, 7, 8, 9

c. **A:** 1, 2, 3, 4, 5; **B:** 6, 7, 8, 9, 10

d. **A:** 0, 10, 50, 60, 100; **B:** 0, 100, 500, 600, 1000

5. **Means and SDs.** For each part, compare distributions A and B based on their means and standard deviations. You do not need to calculate these statistics; simply state how the means and the standard deviations compare. Make sure to explain your reasoning. *Hint:* It may be useful to sketch dot plots of the distributions.

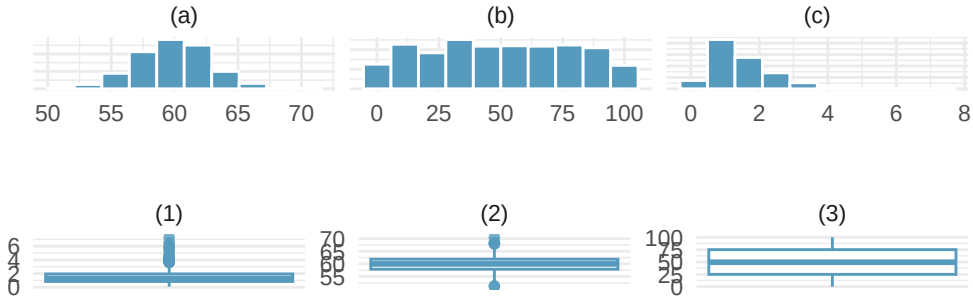
a. **A:** 3, 5, 5, 5, 8, 11, 11, 11, 13; **B:** 3, 5, 5, 5, 8, 11, 11, 11, 20

b. **A:** -20, 0, 0, 0, 15, 25, 30, 30; **B:** -40, 0, 0, 0, 15, 25, 30, 30

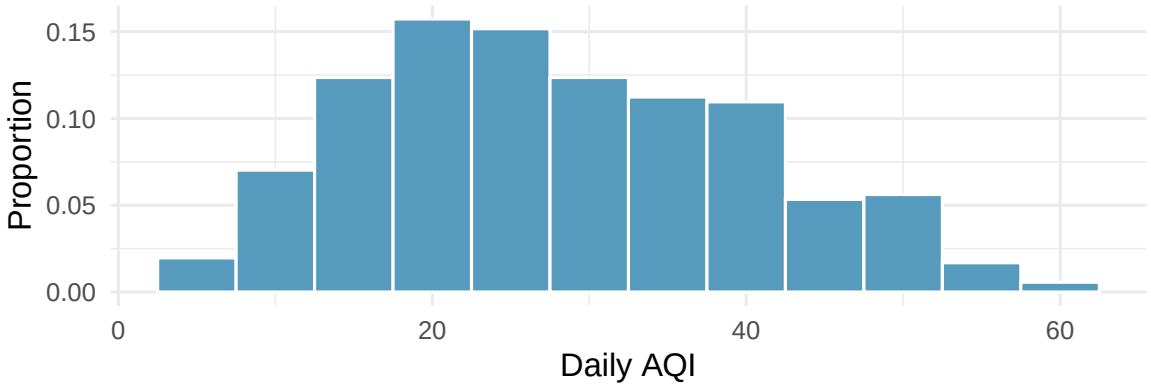
c. **A:** 0, 2, 4, 6, 8, 10; **B:** 20, 22, 24, 26, 28, 30

d. **A:** 100, 200, 300, 400, 500; **B:** 0, 50, 300, 550, 600

6. **Histograms and box plots.** Describe (in words) the distribution in the histograms below and match them to the box plots.


(a)
50
55
60
65
70
(b)
0
25
50
75
100
(c)
0
2
4
6
8
(1)
(2)
(3)

7. **Air quality.** Daily air quality is measured by the air quality index (AQI) reported by the Environmental Protection Agency. This index reports the pollution level and what associated health effects might be a concern. The index is calculated for five major air pollutants regulated by the Clean Air Act and takes values from 0 to 300, where a higher value indicates lower air quality. AQI was reported for a 356 days in 2022 in Durham, NC. The histogram below shows the distribution of the AQI values on these days.



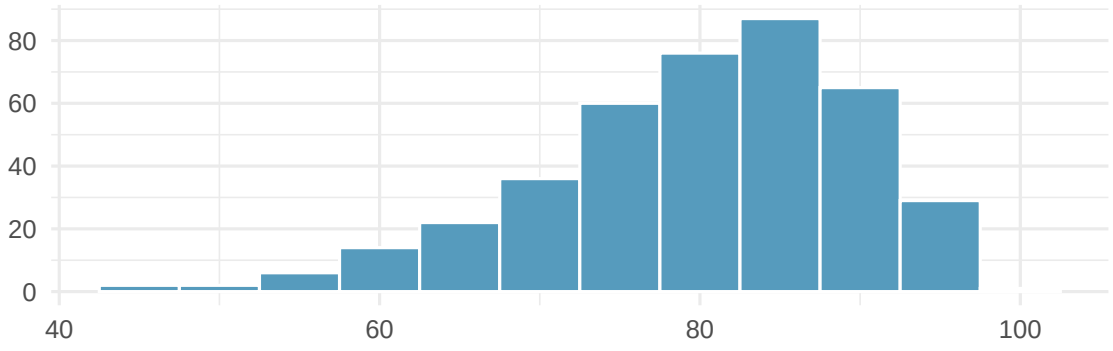
a. Estimate the median AQI value of this sample.

b. Would you expect the mean AQI value of this sample to be higher or lower than the median? Explain your reasoning.

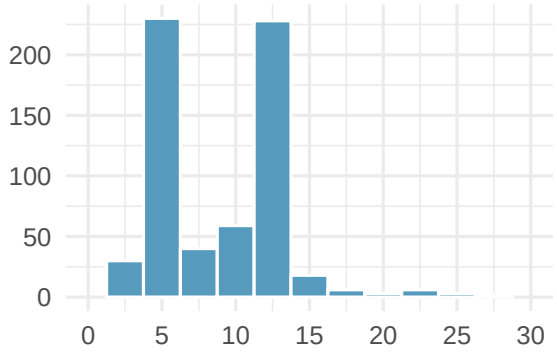
c. Estimate Q1, Q3, and IQR for the distribution.

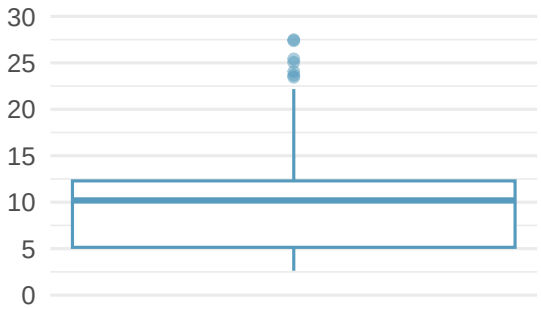
d. Would any of the days in this sample be considered to have an unusually low or high AQI? Explain your reasoning.

8. **Median vs. mean.** Estimate the median for the 400 observations shown in the histogram, and note whether you expect the mean to be higher or lower than the median.



9. **Histograms vs. box plots.** Compare the two plots below. What characteristics of the distribution are apparent in the histogram and not in the box plot? What characteristics are apparent in the box plot but not in the histogram?


200
150
100
50
0
0
5
10
15
20
25
30


30
25
20
15
10
5
0

10. **Facebook friends.** Facebook data indicate that 50% of Facebook users have 100 or more friends, and that the average friend count of users is 190. What do these findings suggest about the shape of the distribution of number of friends of Facebook users? (Backstrom 2011)

11. **Distributions and appropriate statistics.** For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

a. Number of pets per household.

b. Distance to work, i.e., number of miles between work and home.

c. Heights of adult males.

d. Age at death.

e. Exam grade on an easy test.

12. **Distributions and appropriate statistics.** For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

a. Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000, and there are a meaningful number of houses that cost more than $6,000,000.

b. Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000, and very few houses that cost more than $1,200,000.

c. Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

d. Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than all the other employees.

e. Gestation time in humans where 25% of the babies are born by 38 weeks of gestation, 50% of the babies are born by 39 weeks, 75% of the babies are born by 40 weeks, and the maximum gestation length is 46 weeks.

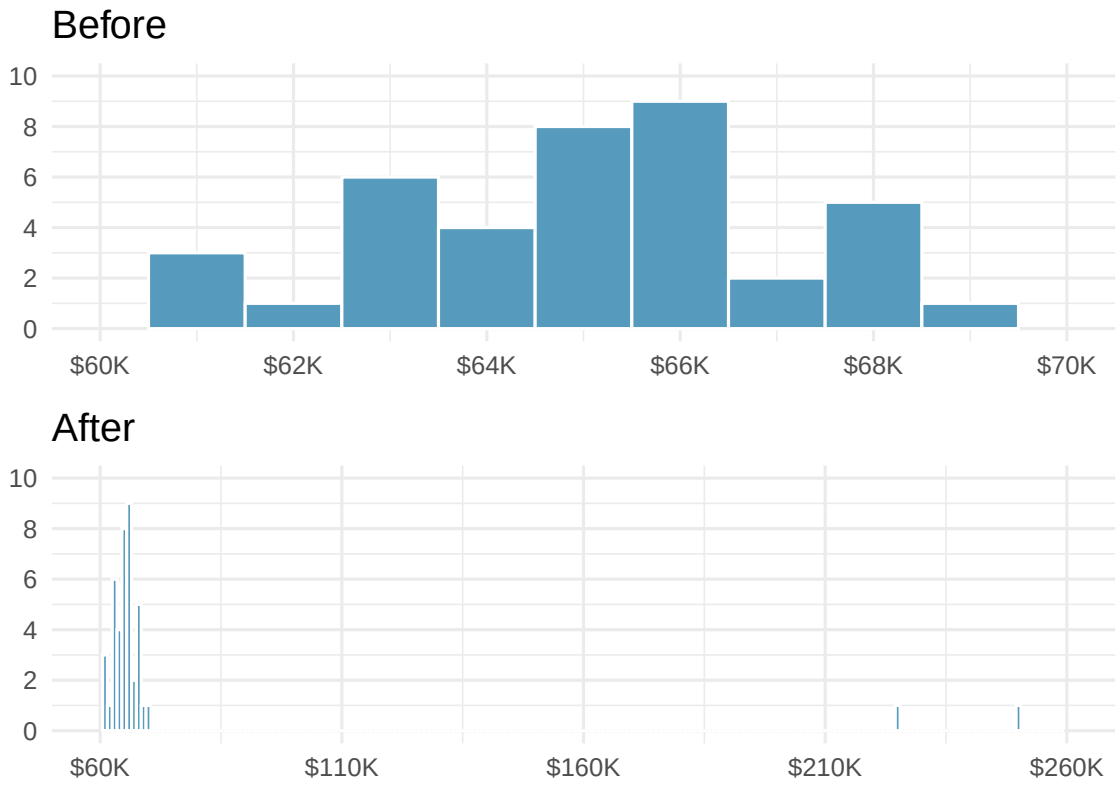
13. **TV watchers.** College students in a statistics class were asked how many hours of television they watch per week, including online streaming services. This sample yielded an average of 8.28 hours, with a standard deviation of 7.18 hours. Is the distribution of number of hours students watch television weekly symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.

14. **Exam scores.** The average on a history exam (scored out of 100 points) was 85, with a standard deviation of 15. Is the distribution of the scores on this exam symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.

15. **Midrange.** The *midrange* of a distribution is defined as the average of the maximum and the minimum of that distribution. Is this statistic robust to outliers and extreme skew? Explain your reasoning.

16. **Stats scores.** The final exam scores of twenty introductory statistics students, arranged in ascending order, as as follows: 57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94. Suppose students who score above the 75th percentile on the final exam get an A in the class. How many students will get an A in this class?

17. **Income at the coffee shop.** The first histogram below shows the distribution of the yearly incomes of 40 patrons at a college coffee shop. Suppose two new people walk into the coffee shop: one making $225,000 and the other $250,000. The second histogram shows the new income distribution. Summary statistics are also provided, rounded to the nearest whole number.



	n	Min	Q1	Median	Mean	Max	SD
Before	40	$60,679	$60,818	$65,238	$65,089	$69,885	$2,122
After	42	$60,679	$60,838	$65,352	$73,299	$250,000	$37,321

a. Would the mean or the median best represent what we might think of as a typical income for the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

b. Would the standard deviation or the IQR best represent the amount of variability in the incomes of the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

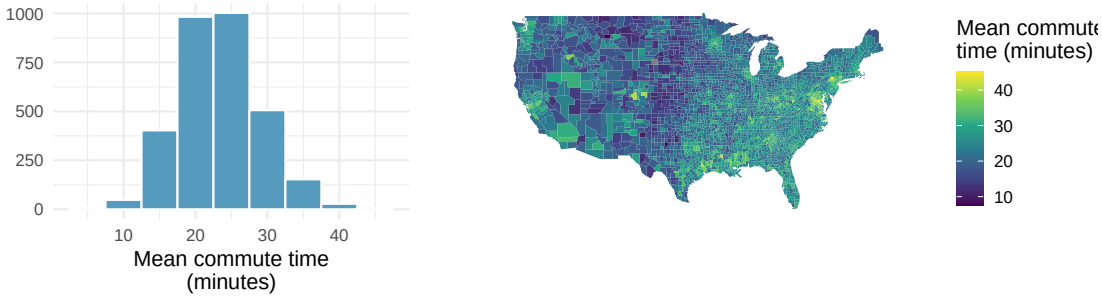
18. **A new statistic.** The statistic $\frac{\bar{x}}{median}$ can be used as a measure of skewness. Suppose we have a distribution where all observations are greater than 0, $x_i > 0$. What is the expected shape of the distribution under the following conditions? Explain your reasoning.

a. $\frac{\bar{x}}{median} = 1$

b. $\frac{\bar{x}}{median} < 1$

c. $\frac{\bar{x}}{median} > 1$

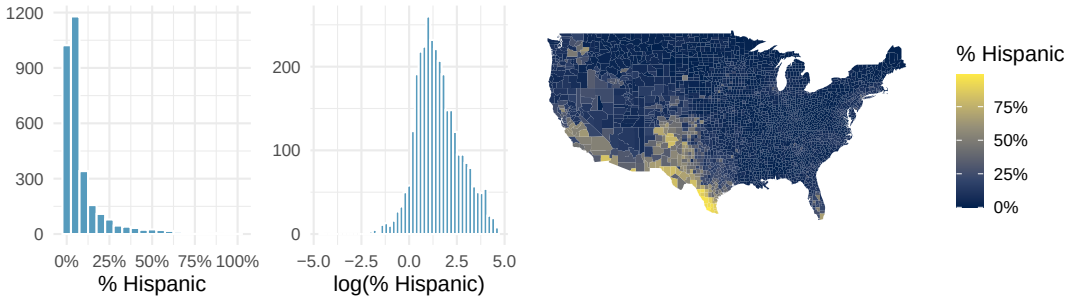
19. **Commute times.** The US census collects data on the time it takes Americans to commute to work, among many other variables. The histogram below shows the distribution of mean commute times in 3,142 US counties in 2017. Also shown below is a spatial intensity map of the same data.



a. Describe the numerical distribution and comment on whether a log transformation may be advisable for these data.

b. Describe the spatial distribution of commuting times using the map.

20. **Hispanic population.** The US census collects data on race and ethnicity of Americans, among many other variables. The histogram below shows the distribution of the percentage of the population that is Hispanic in 3,142 counties in the US in 2010. Also shown is a histogram of logs of these values.



a. Describe the numerical distribution and comment on why we might want to use log-transformed values in analyzing or modeling these data.

b. What features of the distribution of the Hispanic population in US counties are apparent in the map but not in the histogram? What features are apparent in the histogram but not the map?

c. Is one visualization more appropriate or helpful than the other? Explain your reasoning.

21. **Empirical rule for adult IQ scores.** Adult IQ scores are approximately bell-shaped with mean $\mu = 100$ and standard deviation $\sigma = 15$.

a. According to the empirical rule, about 68% of adults have IQ scores in what range?

b. About 95% of adults have IQ scores in what range?

c. About 99.7% of adults have IQ scores in what range?

d. A common cutoff for "gifted" is IQ ≥ 130. Roughly what percent of adults fall at or above this cutoff? (Hint: 130 is exactly two SDs above the mean, and the empirical rule tells you about the two-tailed range.)

22. **When the empirical rule fails.** A dataset of annual household incomes in a large city has a mean of \$68,000 and a standard deviation of \$54,000. A student uses the empirical rule to claim that about 95% of households in this city earn between $\$68{,}000 - 2(\$54{,}000) = -\$40{,}000$ and $\$68{,}000 + 2(\$54{,}000) = \$176{,}000$.

a. What's obviously wrong with the lower endpoint of the interval? What does this suggest about the distribution's shape?

b. Why is the empirical rule inappropriate for this dataset even at the upper end?

c. What shape would you expect a distribution of household incomes to have? Sketch or describe it in words.

d. For a distribution shaped like this one, would the mean or the median be a better summary of a "typical" household income? Why?

Datasets: `cia_factbook` (openintro) | `pm25_2022_durham` (openintro) | `county_complete` (usdata)

StatLens Exercises

The following exercises use StatLens, an interactive statistics tool. Each exercise includes a link that opens a pre-loaded dataset — explore, interact, and answer the questions.

1. **Does binning change the story?** Open the US county poverty rates in StatLens. This dataset contains poverty rates (%) for 3,137 US counties.
   a. Describe the shape of the distribution with the default number of bins. Is it symmetric, left-skewed, or right-skewed? Is it unimodal?
   b. Use the **Bins** control to reduce the number of bins to about 5–6. Does your conclusion about the shape change? What detail is lost?
   c. Now increase the bins to 30–40. Does your shape conclusion change? What new detail (if any) becomes visible?
   d. Based on your exploration, explain in your own words why two histograms of the same data can look somewhat different yet tell the same story.

2. **What do box plots hide?** Open the email character counts in StatLens. This dataset records the number of characters (in thousands) in 50 emails. Numeric labels are hidden so you can focus on visual patterns.

   a. View the data as a **histogram**. Describe the shape. Are there any gaps or clusters that stand out?

   b. Switch to a **dotplot**. With only 50 observations, can you see individual emails? Identify approximately how many emails have more than 25,000 characters.

   c. Now switch to a **boxplot**. Can you still tell whether the distribution has one peak or two? Can you identify the same gaps or clusters you saw in the histogram?

   d. A classmate claims "box plots and histograms show the same information, just in different ways." Do you agree? Use what you observed to explain why or why not.

3. **Choosing the right chart for the question.** Open the US county population data in StatLens. This dataset contains the 2017 population of 3,137 US counties.

   For each question below, state which chart type (**histogram**, **dotplot**, or **boxplot**) would be most useful, then switch to that chart type to answer it.

   a. What is the approximate median county population? Which chart type shows this most directly?

   b. Are there counties with populations over 5 million? How many? Which chart type makes this easiest to determine?

   c. Is the distribution symmetric, left-skewed, or right-skewed? Which chart type makes the skewness most visually obvious?

   d. Try increasing the number of histogram bins to 50 or more. What happens? Why does this dataset pose a challenge for histograms?

4. **Mean vs. median: which tells the truth?** Open the NBA player heights in StatLens, then answer parts (a)–(b). Next, open the Ames home sale prices and answer parts (c)–(d).

   a. For the NBA heights data, view the **histogram** and describe the shape. Now look at the **mean** and **median** in the summary statistics panel. Are they close together or far apart?

   b. Would it matter much whether you reported the mean or the median as the "typical" NBA player height? Why or why not?

   c. For the Ames home prices, view the **histogram** and describe the shape. Look at the mean and median. How far apart are they? Which is larger?

   d. A real estate agent wants to tell a client what a "typical" home costs in Ames. Should the agent report the mean or the median? Explain, using the shape of the distribution to support your answer.

5. **Checking the empirical rule.** Open the NBA player heights in StatLens. Heights of professional basketball players are approximately bell-shaped.
   a. Record the **mean** and **standard deviation** from the summary statistics panel.
   b. Compute the interval for "within 1 standard deviation of the mean": $(\bar{x} - s,\ \bar{x} + s)$. According to the empirical rule, about what percentage of players should have heights in this range?
   c. Look at the histogram. Estimate (roughly) whether about 68% of the data appears to fall within the interval you computed. Does the empirical rule seem to hold?
   d. Now compute the interval for "within 2 standard deviations": $(\bar{x} - 2s,\ \bar{x} + 2s)$. According to the empirical rule, about what percentage should fall in this range? Looking at the histogram, does nearly all the data fall within this interval?
   e. Would you expect the empirical rule to work as well for the Ames home price data? Why or why not? (Hint: think about the shape.)

Chapter 5

Relationships Between Variables

The preceding chapters focused on exploring one variable at a time. But the most interesting statistical questions involve **relationships between variables**. Does higher education lead to higher income? Is a drug more effective for one group than another? This chapter introduces tools for exploring relationships: side-by-side box plots and faceted histograms for comparing a quantitative variable across groups, and scatterplots for examining the association between two quantitative variables. We close with the critical distinction between association and causation.

5.1 Comparing groups

Some of the most interesting investigations involve comparing a quantitative variable across groups defined by a categorical variable.

5.1.1 Side-by-side box plots

Side-by-side box plots place a box plot for each group on the same scale, making it easy to compare centers, spreads, and shapes.

Consider comparing median household income for US counties, grouped by whether the county experienced population growth from 2010 to 2017:

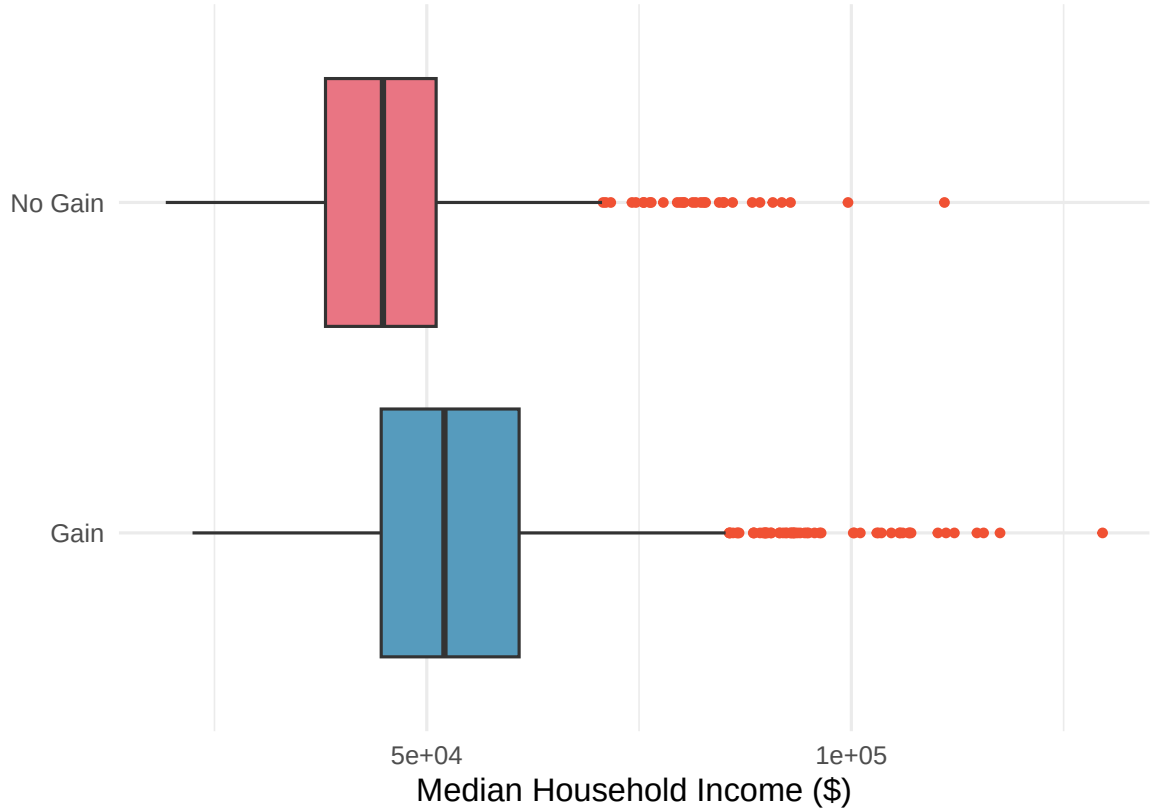


Figure 5.1: Side-by-side box plots of median household income by population change status (gain vs. no gain). Counties with population gain have a higher median income. Open in StatLens

Using the box plots, compare the income distributions for the two groups. What do you notice about the centers, spreads, and shapes?

Show answer

Counties with population gains tend to have higher income (median about $52,000) compared to counties without gains (median about $45,000). The variability is also slightly larger for the gain group, as seen in the wider IQR. Both distributions show slight to moderate right skew, with outliers on the high end.

When are side-by-side box plots particularly useful? When might histograms be better?

Side-by-side box plots are excellent for comparing centers (medians) and spreads (IQRs) across groups in a compact display. However, they cannot reveal details about shape such as bimodality. Histograms or density plots are better when you want to see the full shape of each distribution, including the number of modes and the exact nature of the skew.

5.1.2 Faceted histograms

Faceting splits a graphical display into panels based on the levels of a categorical variable. Each panel shows the distribution for one group, using the same scale for easy comparison.

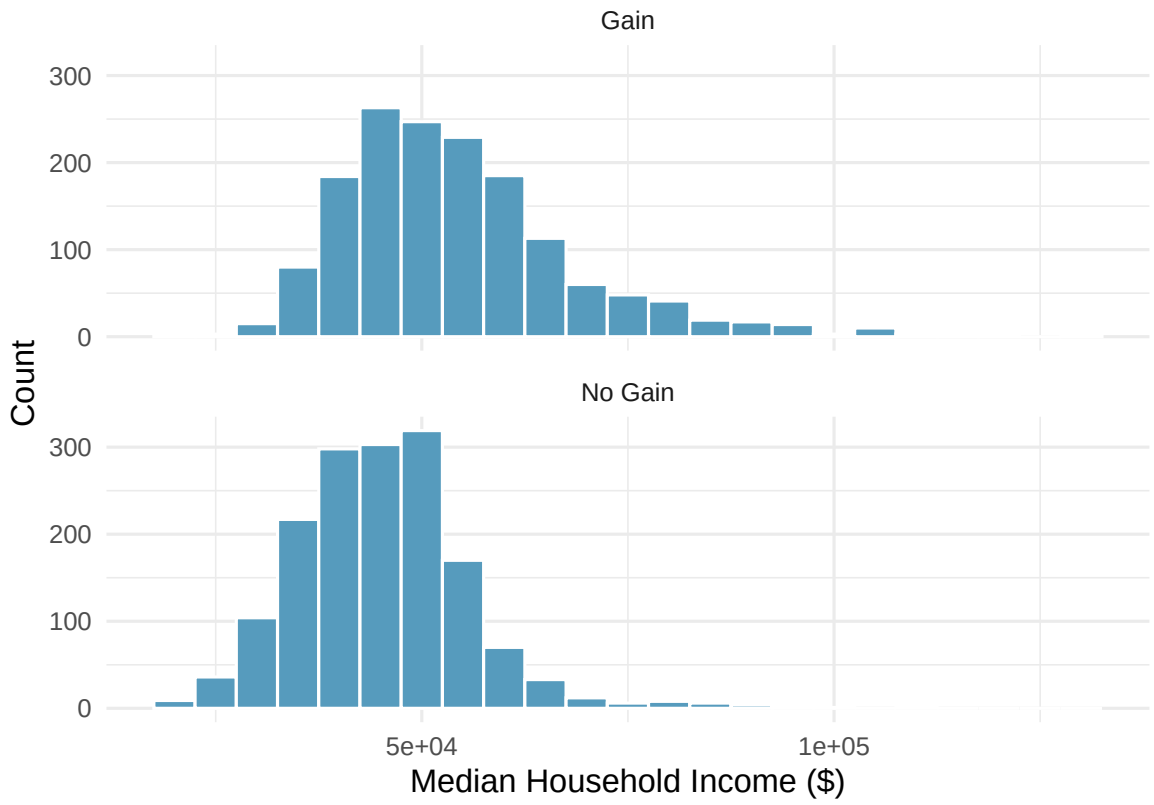


Figure 5.2: Faceted histograms of median household income, separated into panels for counties with population gain and those with no gain. Open in StatLens

Faceting extends naturally to multiple categorical variables. For example, we could split the income data by both population change *and* whether the county is in a metropolitan area, creating a 2-by-2 grid of histograms.

What are the advantages of faceted histograms compared to side-by-side box plots?

Faceted histograms reveal the full shape of each distribution, including modality (unimodal vs. bimodal), the exact nature of any skewness, and gaps or clusters. Box plots, on the other hand, are more compact and make it easier to compare medians and IQRs at a glance, but they cannot show features like bimodality. The two displays complement each other well.

5.1.3 Ridge plots

A **ridge plot** (also called a joy plot) is a set of overlapping density plots, one for each group, drawn on the same horizontal scale but offset vertically. Ridge plots are useful for comparing the shapes of many distributions at once while maintaining a compact display. With only two groups, as in the figure below, the density curves can simply be overlaid on a shared baseline instead of being offset.

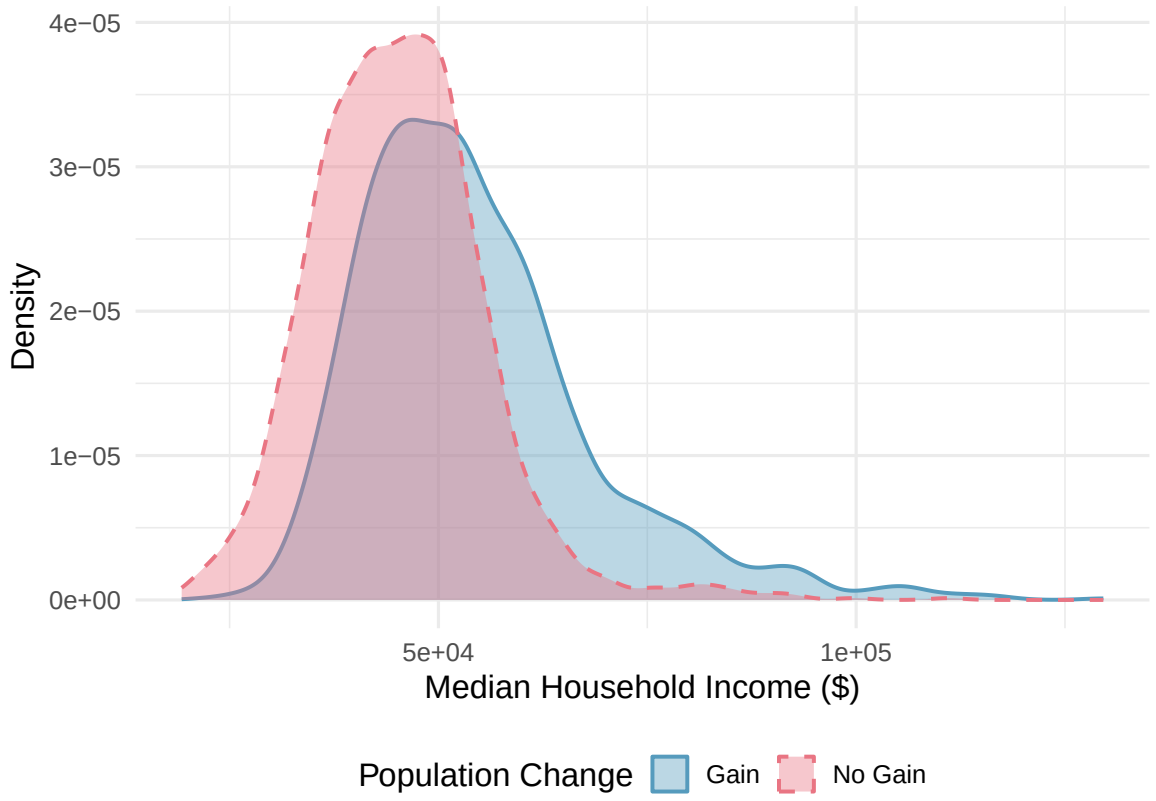


Figure 5.3: Overlaid density curves comparing median household income for counties with and without population gain. Each group gets its own smoothed density curve; a ridge plot would stack these curves with a vertical offset. Open in StatLens

What does a ridge plot reveal that side-by-side box plots do not?

Show answer

Ridge plots show the full shape of each distribution, including modality and skewness, similar to histograms. Box plots compress all of this shape information into a five-number summary. Ridge plots are especially useful when you want to compare shapes across many groups in a compact space.

See it in action. Open the Comparing Distributions Across Groups activity to walk through side-by-side box plots, histograms, and density overlays on birth weight grouped by maternal smoking status.

5.2 Scatterplots

5.2.1 Visualizing two quantitative variables

A **scatterplot** displays the relationship between two quantitative variables. Each observation is represented as a point, with one variable on the horizontal axis and the other on the vertical axis.

A **scatterplot** provides a case-by-case view of data for two quantitative variables. Each point represents one observation. The horizontal axis shows the values of one variable and the vertical axis shows the values of the other.

Consider a dataset of 50 loans, where we plot total income on the horizontal axis and loan amount on the vertical axis:

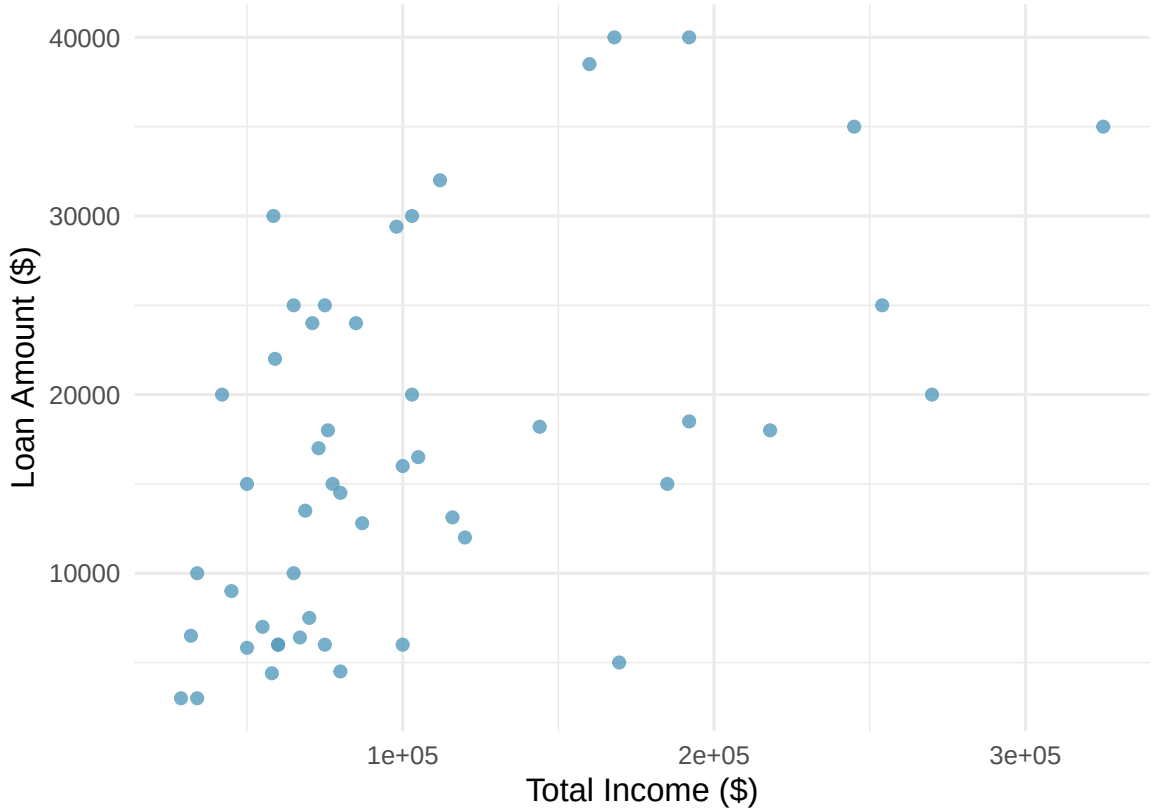


Figure 5.4: A scatterplot of loan amount versus total income for 50 loans. There is a moderate positive association. Open in StatLens

What do scatterplots reveal about the data, and how are they useful?

Show answer

Scatterplots are helpful for quickly spotting associations between two quantitative variables — whether they tend to increase together, decrease together, or show no pattern. They can also reveal whether the relationship is linear or curved, how strong the association is, and whether there are any unusual points.

5.2.2 Describing associations

When examining a scatterplot, we describe the association using four characteristics:

Describing the association between two quantitative variables:

1. **Direction**: Is the association **positive** (both variables tend to increase together) or **negative** (one increases as the other decreases)?
2. **Form**: Is the relationship approximately **linear** (following a straight-line pattern) or **nonlinear** (curved)?
3. **Strength**: Is the association **strong** (points closely follow a pattern), **moderate**, or **weak** (points are loosely scattered)?

4. **Unusual observations**: Are there any **outliers** — points that deviate markedly from the overall pattern?

A scatterplot shows the poverty rate vs. median household income for over 3,000 US counties. How would you describe the association?

The association is **negative** (as poverty rate increases, median household income tends to decrease), **nonlinear** (the relationship curves, with income decreasing more steeply at lower poverty rates and leveling off at higher poverty rates), and **moderately strong** (the points follow the curving pattern fairly closely, though there is noticeable scatter). There do not appear to be any extreme outliers, though counties with very high poverty rates or very high incomes stand out somewhat.

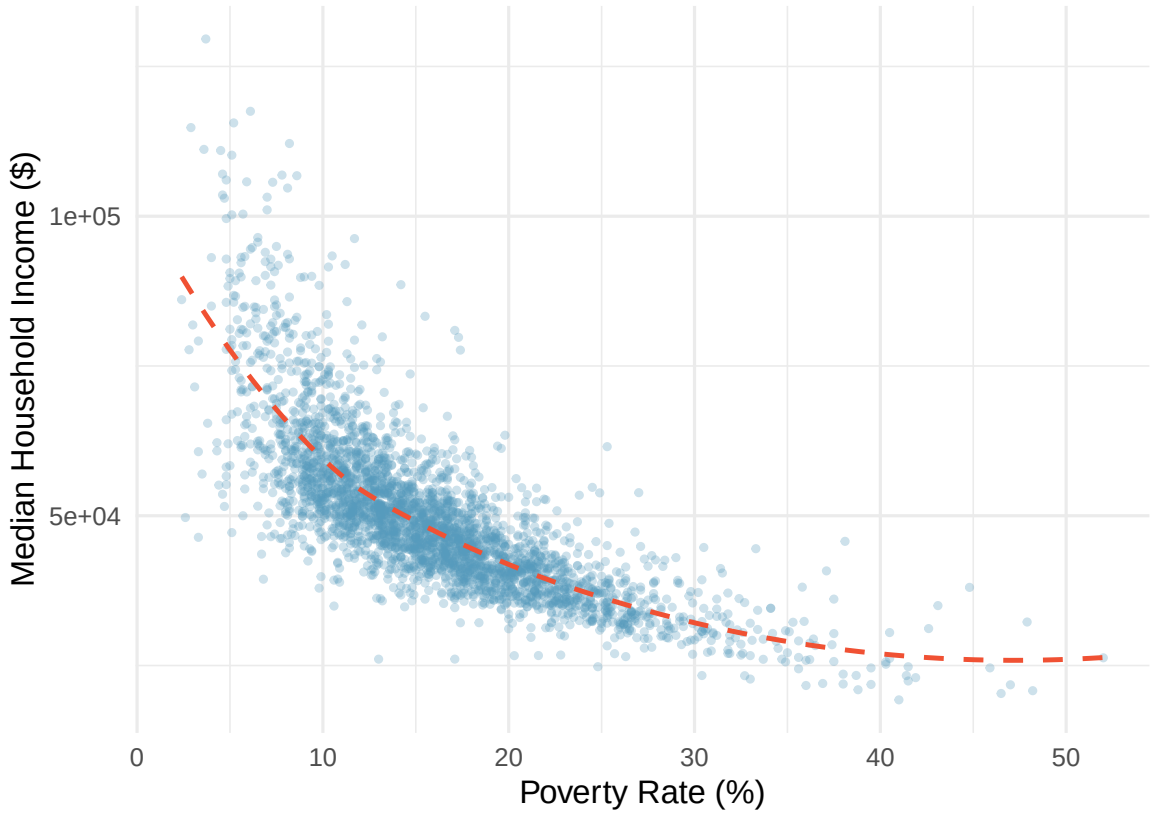


Figure 5.5: A scatterplot of median household income versus poverty rate for US counties, with a curved trend line. The relationship is negative and nonlinear. Open in StatLens

Describe two variables that would have a horseshoe-shaped (∩) association in a scatterplot.

Show answer

Consider a situation where the vertical axis represents something "good" and the horizontal axis represents something that is good only in moderation. For example, health and water consumption: we need some water to survive, but consuming too much becomes dangerous. Athletic performance and practice time is another example — some practice improves performance, but overtraining leads to injury and decline.

For the 50-loan dataset, the scatterplot of total income versus loan amount shows that borrowers with

higher incomes tend to take out larger loans. Describe this association.

The association is **positive** (higher income is associated with higher loan amounts), approximately **linear** (the general trend follows a straight line, though the pattern is loose), and **weak to moderate** (there is considerable scatter — many borrowers with similar incomes take out very different loan amounts). There are a few borrowers with income above $250,000 who stand out on the right side of the plot.

See it in action. Open the Describing Associations in Scatterplots activity to practice identifying direction, form, strength, and unusual observations on possum body measurements.

5.2.3 Scatterplots with a third variable

We can add a third variable to a scatterplot by using color, shape, or size to represent a categorical or quantitative variable. For example, in a scatterplot of income vs. poverty rate, we could color the points by whether the county is in a metropolitan area. This technique helps reveal whether the relationship between two quantitative variables differs across groups.

5.3 Association vs. causation

Throughout this chapter, we have explored associations between variables. But association is not the same as causation.

Association does not imply causation.

When two variables are associated (correlated), it means they tend to vary together — but that does not necessarily mean one *causes* the other. There may be a **confounding variable** (a third variable) that is related to both, creating the illusion of a direct relationship.

A study finds a strong positive association between ice cream sales and drowning deaths. Does eating ice cream cause drowning?

Of course not. The association exists because both ice cream sales and drowning are related to a confounding variable: **temperature** (or season). In hot weather, people buy more ice cream *and* more people go swimming, which increases the risk of drowning. Temperature is the lurking variable that drives both.

A **confounding variable** (also called a **lurking variable**) is a variable that is related to both the explanatory and response variables. Confounding variables can create misleading associations between two variables that have no direct causal relationship.

A study finds that students who eat breakfast tend to have higher GPAs than students who skip breakfast. Can we conclude that eating breakfast causes higher GPAs?

Show answer

Not from this information alone. This is likely an observational study, and there could be confounding variables. For example, students who eat breakfast may come from more supportive home environments, may have better time management skills, or may generally take better care of their health — all of which could independently contribute to higher GPAs. To establish a causal connection, we would need a randomized experiment.

5.3.1 When can we infer causation?

Establishing a causal relationship between two variables generally requires:

1. **A randomized experiment**: Randomly assigning subjects to treatment and control groups helps eliminate confounding variables. If a difference is observed between groups, we can more confidently attribute it to the treatment.
2. **Temporal precedence**: The cause must precede the effect in time.
3. **Ruling out confounders**: We must be able to argue that no lurking variables explain the observed association.

In observational studies (where we observe but do not control the variables), we can identify associations but generally cannot draw causal conclusions. This is one of the most important distinctions in statistics.

Be especially skeptical of causal claims from observational data. Media reports often say "X causes Y" when the underlying study only found an association. Always ask: Was this a randomized experiment? Could there be confounding variables?

5.4 Chapter review

5.4.1 Summary

This chapter introduced tools for exploring relationships between variables. For **comparing a quantitative variable across groups** defined by a categorical variable, side-by-side box plots provide a compact comparison of centers and spreads, faceted histograms reveal full distributional shapes, and ridge plots overlay density curves for compact multi-group comparison. For **two quantitative variables**, scatterplots reveal the direction, form, strength, and unusual features of an association. We closed with the fundamental principle that **association does not imply causation** — confounding variables can create misleading associations, and establishing causation generally requires a randomized experiment.

5.4.2 Key terms

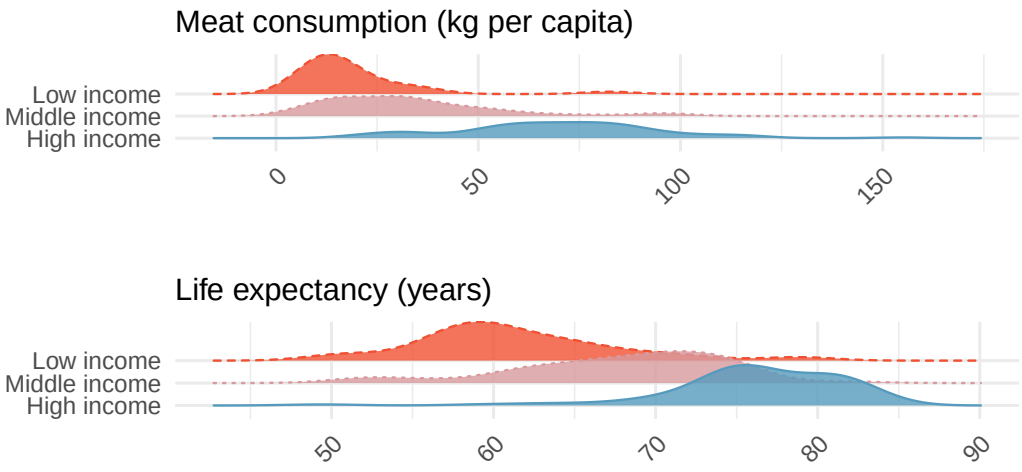
- Side-by-side box plot, faceted plot, ridge plot
- Scatterplot
- Positive association, negative association
- Linear, nonlinear
- Direction, form, strength (of an association)
- Confounding variable (lurking variable)
- Association vs. causation

5.5 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Meat consumption and life expectancy.** In data collected for You et al. (2022), total meat intake is associated with life expectancy (at birth) in 175 countries. Meat intake is measured in kg per capita per year (averaged over 2011 to 2013). The two ridge plots show an association between income and meat consumption (higher income countries tend to eat more meat) and

an association between income and life expectancy (higher income countries have higher life expectancy).



a. Do the graphs above demonstrate that meat consumption and life expectancy are associated? That is, can you tell if countries with low meat consumption have low life expectancy? Explain.

b. Let's assume that you had a plot comparing meat consumption and life expectancy, and they *do* seem associated. Your friend says that the plot shows that high meat consumption leads to a longer life. You correctly say, no, we can't tell if there is a causal realtionship because the relationship is confounded by income level. Explain what you mean.

c. How can you investigate the relationship between meat consumption and life expectancy in the presence of confounding variables (like income)?

2. **Florence Nightingale.** Florence Nightingale was a nurse in the Crimean War and an early statistician. In her notes, she opined, "In comparing the deaths of one hospital with those of another, any statistics are justly considered absolutely valueless which do not give the ages, the sexes, and the diseases of all the cases." (Nightingale 1859)

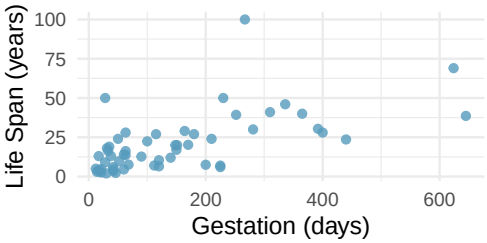
a. Nightingale describes three confounding variables to consider when comparing death rates across hospitals. What are they? Describe what makes each variable potentially confounding.

b. Provide two additional potential confounding variables for this situation. Check to make sure that the variables are associated with both the explanatory variable (hospital) and the response variable (death).

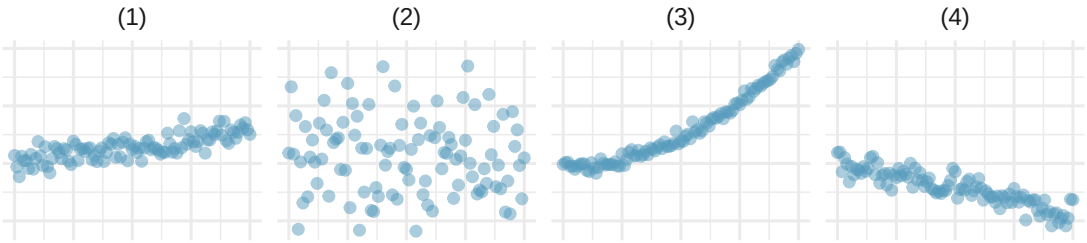
c. Why does Nightingale say that the statistics are "valueless" if given without being broken down by age, sex, and disease? Explain.

3. **Mammal life spans.** Data were collected on life spans (in years) and gestation lengths (in days) for 62 mammals. A scatterplot of life span versus length of gestation is shown below. (Allison and Cicchetti 1975)

    a. What type of an association is apparent between life span and length of gestation?
    b. What type of an association would you expect to see if the axes of the plot were reversed, i.e., if we plotted length of gestation versus life span?
    c. Are life span and length of gestation independent? Explain your reasoning.



4. **Associations.** Indicate which of the plots show (a) a positive association, (b) a negative association, or (c) no association. Also determine if the positive and negative associations are linear or nonlinear. Each part may refer to more than one plot.

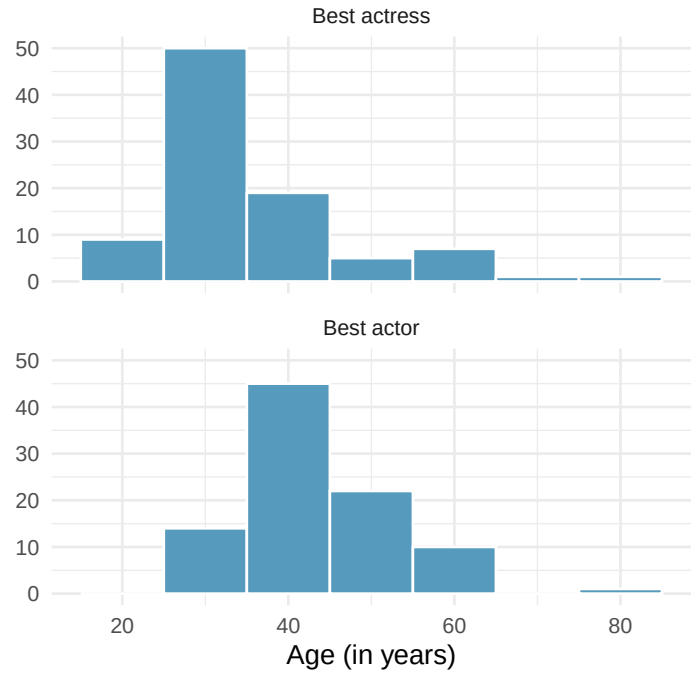


5. **Reproducing bacteria.** Suppose that there is only sufficient space and nutrients to support one million bacterial cells in a petri dish. You place a few bacterial cells in this petri dish, allow them to reproduce freely, and record the number of bacterial cells in the dish over time. Sketch a plot representing the relationship between number of bacterial cells and time.

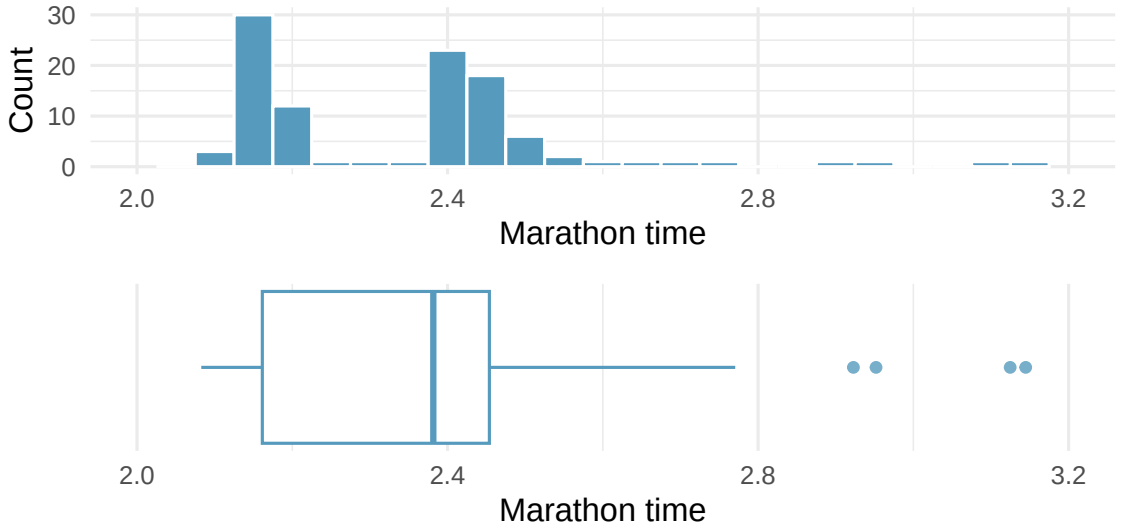
6. **Office productivity.** Office productivity is relatively low when the employees feel no stress about their work or job security. However, high levels of stress can also lead to reduced employee productivity. Sketch a plot to represent the relationship between stress and productivity.

7. **Oscar winners.** The first Oscar awards for best actor and best actress were given out in 1929. The histograms below show the age distribution for all of the best actor and best actress winners from 1929 to 2019. Summary statistics for these distributions are also provided. Compare the distributions of ages of best actor and actress winners.

8. **NYC marathon winners.** The histogram and box plots below show the distribution of finishing times for male and female (combined) winners of the New York City Marathon between 1970 and 2023.



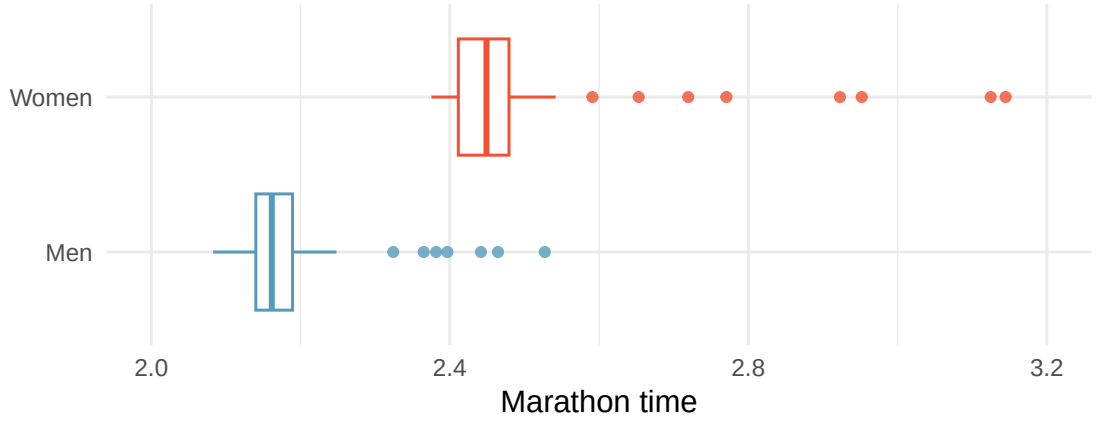
	Mean	SD	n
Best actor	43.8	8.8	92
Best actress	36.2	11.9	92



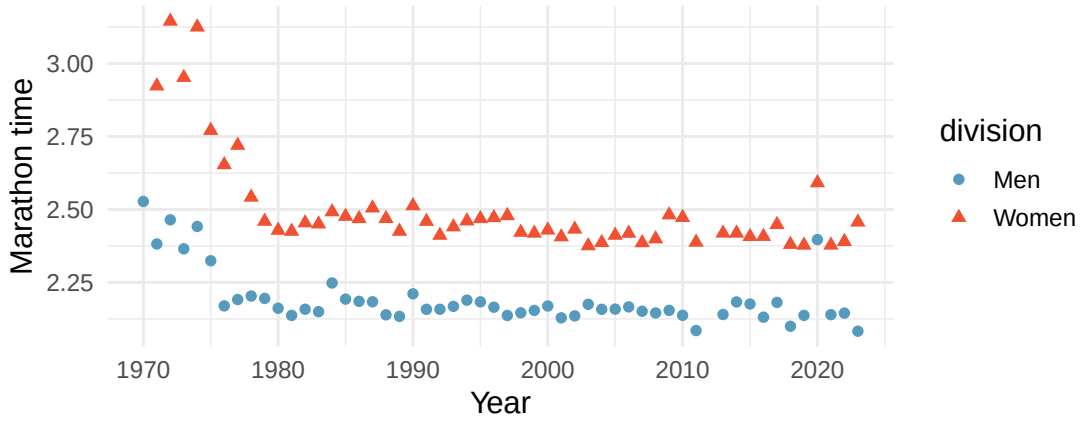
a. What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

b. What may be the reason for the bimodal distribution? Explain.

c. Compare the distribution of marathon times for men and women based on the box plot shown below.



d. The time series plot shown below is another way to look at these data. Describe what is visible in this plot but not in the others.



Datasets: `mammals` (openintro) | `oscars` (openintro) | `nyc_marathon` (openintro)

StatLens Exercises

The following exercises use StatLens, an interactive statistics tool. Each exercise includes a link that opens a pre-loaded dataset — explore, interact, and answer the questions.

1. **Fuel efficiency by transmission type.** Open the EPA fuel economy data in StatLens. This dataset contains city miles per gallon (MPG) for 201 vehicles from the 2021 EPA fuel economy guide, grouped by transmission type (automatic vs. manual).
   a. Start with the **box plots**. Compare the two groups: which transmission type has the higher median city MPG? Which group has more variability (larger IQR)?
   b. Switch to **Histogram** view. Describe the shape of each group's distribution. Are they symmetric or skewed? Unimodal or bimodal?
   c. Switch to **Density** overlay. Where do the two curves overlap the most? Based on this overlap, would you say the difference between transmission types is large relative to the variability within each group?
   d. Using the summary statistics table, write a complete comparison of the two groups that includes shape, center, spread, and any unusual features.

2. **Starbucks nutrition: calories and protein.** Open the Starbucks nutrition data in StatLens. This dataset contains calorie and protein information for Starbucks food menu items.

   a. Describe the **direction** of the association between calories and protein. Is it positive, negative, or neither?

   b. Describe the **form**. Does the relationship appear to be linear or nonlinear?

   c. Describe the **strength**. Are the points tightly clustered around a trend, or widely scattered?

   d. Are there any **unusual observations** — menu items that have much more (or less) protein than you would expect for their calorie count? Identify approximately where they fall on the plot.

   e. Write a one-sentence summary that describes the association using all four characteristics (direction, form, strength, unusual observations).

Explore — Unit Review

Unit 1 is about *looking before you leap*: understanding what your data are, how they were collected, and what an honest summary looks like — long before any inference. This review ties the five chapters together and drills the central skill: **matching a question to the right display and summary.**

The big picture

Exploring data is a short, repeatable workflow:

1. **Know your data** — cases and variables, and each variable's **type** (categorical or numerical). The type decides everything that follows.
2. **Know where it came from** — study design. *How* the data were collected sets the ceiling on what you can conclude: an **experiment** with random assignment can support a cause-and-effect claim; an **observational** study can show association but is vulnerable to confounding. Random *sampling* governs how far results **generalize**.
3. **Summarize honestly** — one variable at a time (visualizing distributions, describing distributions) and then relationships between two, choosing displays and numbers that fit the variable types and don't mislead.

Which display and which summary?

This single table is the heart of Unit 1. The variables you have — how many, and what type — pick the row:

What you have	Typical display	Numerical summary
One categorical	bar chart, pie/waffle	counts, **proportions**
One numerical	histogram, dotplot, **boxplot**	**center** (mean/median), **spread** (SD/IQR), shape
Two categorical	two-way table, stacked/side-by-side bars	**conditional proportions**
Categorical + numerical	side-by-side boxplots	group means/medians
Two numerical	**scatterplot**	**correlation** r, regression line

Mean vs. median, SD vs. IQR. For roughly symmetric data, the **mean** and **standard deviation** describe center and spread well. When the distribution is **skewed or has outliers**, prefer the **median** and **IQR** — they are *resistant* to extreme values.

A quick decision guide

1. **How many variables** are in the question — one or two?
2. **What type** is each — categorical or numerical?
3. Read the matching row above for the display, then pick the summary; if a single numerical variable is **skewed**, switch from mean/SD to median/IQR.
4. Separately ask the *design* question: **experiment or observational?** (causation vs. association) and **how was the sample drawn?** (generalizability).

Key ideas to carry forward

Case · variable · categorical vs. numerical · explanatory vs. response · observational study vs. experiment · random assignment (→ causation) · **random sampling** (→ generalizability) · **confounding · distribution · center / spread / shape · skew · outlier · conditional proportion · correlation**.

Common pitfalls

- **Association is not causation** — only a randomized **experiment** licenses a cause-and-effect claim.
- **A representative sample is about *how it was drawn*, not how big it is** — a huge convenience sample can still be biased.
- **Pie charts hide comparisons** — bar charts are almost always easier to read.
- **The mean is pulled toward the tail** — in skewed data it can misrepresent a "typical" value; report the median.
- **Correlation describes only *linear* association** — a strong curved pattern can have r near 0.

Exercises

The Unit 1 skill is **matching a question to the right display and summary**. Answers are provided so you can check your work.

1. **Which display and which summary?** For each scenario, name (i) an appropriate **display** and (ii) an appropriate **numerical summary**.

   a. The distribution of commute times (minutes) for 500 employees.
   b. The breakdown of students by major (six categories).
   c. Whether passing a course is related to attending a review session (each yes/no).
   d. How exam score relates to hours studied.
   e. Comparing salaries across three departments.

   Show answer

	Display	Summary
a	histogram or boxplot	median & IQR (commute times are usually right-skewed)
b	bar chart	counts / proportions per major
c	two-way table, side-by-side or stacked bars	conditional proportions (pass rate by attendance)

	Display	Summary
d	scatterplot	correlation r (and a regression line)
e	side-by-side boxplots	group medians (or means) and spreads

2. **Mean or median? SD or IQR?** A dataset of household incomes is strongly right-skewed with a few very high values. Which measures of center and spread should you report, and why?

   Show answer

   Report the **median** and **IQR**. Both are *resistant* to the extreme high incomes; the mean and SD would be inflated by the long right tail and would overstate a "typical" household's income.

3. **What can you conclude?** A study finds that students who use a tutoring center have higher GPAs than those who don't. Two designs are possible: (A) students *chose* whether to attend (observational); (B) students were *randomly assigned* to attend or not (experiment). For each, state whether a **causal** claim ("tutoring raises GPA") is justified.

   Show answer

   (A) No — observational; motivated students may both attend tutoring *and* earn higher GPAs (confounding), so only **association** is supported. **(B) Yes** — random assignment balances other factors across groups, so a difference can be attributed to the tutoring itself.

4. **Spot the misleading choice.** A report uses a **pie chart** to compare the market share of eight competing brands, several of which are close in size. Why is this a poor choice, and what's better?

   Show answer

   Human eyes judge **lengths** far better than **angles/areas**, so near-equal slices are almost impossible to rank in a pie chart — especially with eight categories. A **bar chart** (ordered by size) makes the comparison immediate.

5. **Scope of inference.** A campus survey is collected by polling students as they leave the gym one afternoon. To what population, if any, do the results generalize?

   Show answer

   This is a **convenience sample** of gym-goers at one time — not a random sample of all students. Results may describe these respondents but **do not reliably generalize** to the whole student body (gym-goers likely differ systematically). Sample size doesn't fix this; the sampling *method* does.

Tech Tutorial: Explore

Tech Tutorial — Explore. A *Tech Tutorial* shows you how to carry out the techniques from this unit in the software your section uses. Pick **one** tool and follow its tab; the question, the data, and the statistical reasoning are identical across all three.

- **StatLens** — nothing to install; runs in your browser. The *fundamentals* path.
- **Jamovi** — a free, menu-driven statistics program.
- **R** — code-based; reproducible and what the wider community uses.

Your choice of tool persists across every tabset in the book, so you only pick once.

The data

We use a small, **synthetic** first-day class survey — 50 students, with two categorical variables and three numerical ones. (It stands in for a real survey; no actual student records are involved.)

Variable	Type	Meaning
`year`	categorical	First-year, Sophomore, Junior, Senior
`housing`	categorical	On-campus or Off-campus
`study_hours`	numerical	hours spent studying in a typical week
`sleep_hours`	numerical	typical nightly sleep
`commute_min`	numerical	one-way commute, in minutes

Get the file: `class_survey.csv`. The Jamovi and StatLens steps below load this exact file.

What you'll do

Five core Explore skills, each in your chosen tool:

1. Get to know the data and classify variables
2. Summarize **one categorical** variable
3. Summarize **one numerical** variable
4. Compare a numerical variable **across groups**
5. Explore the **relationship** between two numerical variables

1. Get to know the data

Before any chart, look at the data and decide each variable's type — *categorical* (labels) or *numerical* (measured amounts). This is a thinking step, not a calculation: no tool decides it for you.

5.5.1 StatLens

1. Open the Descriptive Statistics explorer — the class survey loads automatically.
2. Use the **variable dropdown** to see every column. Numerical variables (`study_hours`, `sleep_hours`, `commute_min`) can be charted here; the categorical ones (`year`, `housing`) you'll explore in steps 2 and 4.

5.5.2 Jamovi

1. Download `class_survey.csv` (link above) and open it: ▫ → **Open** → **This PC**.
2. In the **Data** tab, click each variable and check its **Measure type** — set `year` and `housing` to **Nominal**, and the three counts to **Continuous**.

(*Jamovi screenshots to be added.*)

5.5.3 R

```
library(readr)
library(dplyr)
library(ggplot2)

survey <- read_csv("../datasets/class_survey.csv") |>
  mutate(year = factor(year,
           levels = c("First-year", "Sophomore", "Junior", "Senior")))

glimpse(survey)
```

```
Rows: 50
Columns: 5
$ year        <fct> Sophomore, First-year, Junior, First-year, Senior, Sophomo~
$ housing     <chr> "On-campus", "On-campus", "On-campus", "Off-campus", "Off-~
$ study_hours <dbl> 12.0, 7.8, 15.0, 10.1, 5.1, 7.5, 12.1, 11.8, 7.9, 12.5, 6.~
$ sleep_hours <dbl> 7.6, 7.0, 9.0, 6.5, 6.0, 6.7, 7.2, 9.0, 7.3, 6.8, 6.5, 6.6~
$ commute_min <dbl> 1, 3, 3, 32, 21, 97, 11, 33, 16, 5, 40, 8, 16, 7, 11, 8, 1~
```

---

2. Summarize one categorical variable

For a categorical variable we report a **frequency table** (counts and proportions) and a **bar chart**. We'll use `year`.

5.5.4 StatLens

1. Open the One Categorical Variable explorer.
2. Choose **year**. Read the **frequency table** and the **bar chart**: which class year is most common?

5.5.5 Jamovi

1. **Exploration → Descriptives.** Put `year` in **Variables**.
2. Tick **Frequency tables**. Under **Plots**, turn on **Bar plot**.

(*Jamovi screenshots to be added.*)

5.5.6 R

```
# Frequency table: counts and proportions
survey |>
  count(year) |>
  mutate(proportion = round(n / sum(n), 2))
```

```
# A tibble: 4 x 3
  year           n proportion
  <fct>      <int>      <dbl>
1 First-year    18       0.36
2 Sophomore     14       0.28
3 Junior        10       0.2
4 Senior         8       0.16
```

```
ggplot(survey, aes(x = year)) +
  geom_bar(fill = "#3E5496") +
  labs(x = "Class year", y = "Count") +
  theme_minimal(base_size = 12)
```

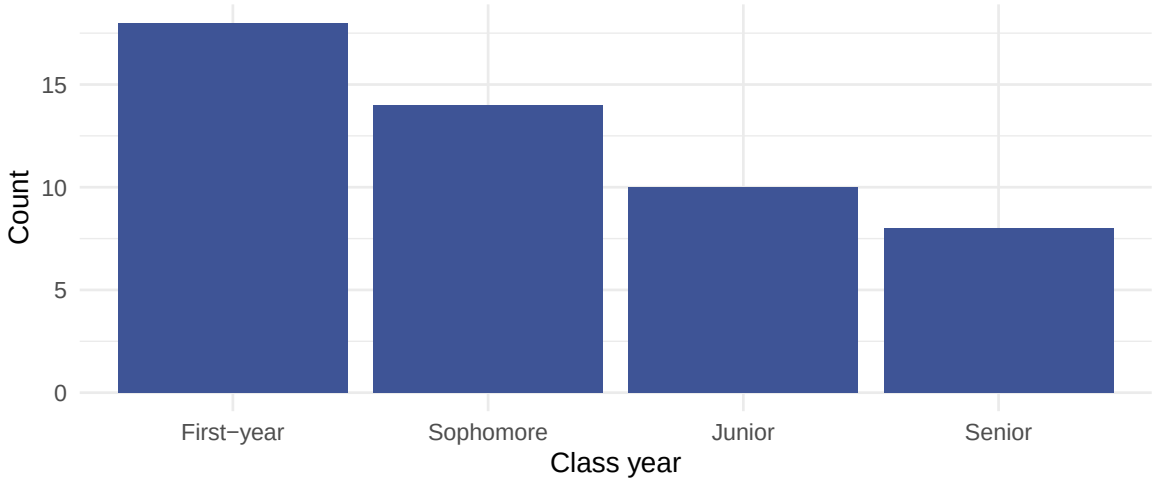


Figure 5.6: Class year of the 50 surveyed students.

---

3. Summarize one numerical variable

For a numerical variable we describe **shape**, **center**, and **spread**, and watch for **outliers**. We'll use `study_hours`. Because the distribution is right-skewed with a high outlier, the **median and IQR** are more trustworthy than the mean and SD.

5.5.7 StatLens

1. Open the Descriptive Statistics explorer.
2. Choose **`study_hours`**. Read the **histogram** (note the right skew and the lone high value), then switch the chart to a **box plot**.
3. Read the **mean, median, SD, and IQR** from the summary panel.

5.5.8 Jamovi

1. **Exploration → Descriptives.** Put `study_hours` in **Variables**.
2. Request **Mean, Median, Std. deviation, IQR**.
3. Under **Plots**, turn on **Histogram** and **Box plot**.

(*Jamovi screenshots to be added.*)

5.5.9 R

```
library(patchwork)

# Mean/SD vs. median/IQR — note the mean sits above the median (skew)
survey |>
  summarize(mean = mean(study_hours), median = median(study_hours),
            sd = sd(study_hours), IQR = IQR(study_hours))
```

```
# A tibble: 1 x 4
   mean median    sd   IQR
  <dbl>  <dbl> <dbl> <dbl>
1  13.9   12.8  6.64  6.67
```

```
p1 <- ggplot(survey, aes(study_hours)) +
  geom_histogram(binwidth = 4, fill = "#3E5496", color = "white") +
  labs(x = "Study hours / week", y = "Count") +
  theme_minimal(base_size = 12)

p2 <- ggplot(survey, aes(y = study_hours)) +
  geom_boxplot(fill = "#3E5496") +
  labs(y = "Study hours / week") +
  theme_minimal(base_size = 12) +
  theme(axis.text.x = element_blank())

p1 + p2
```

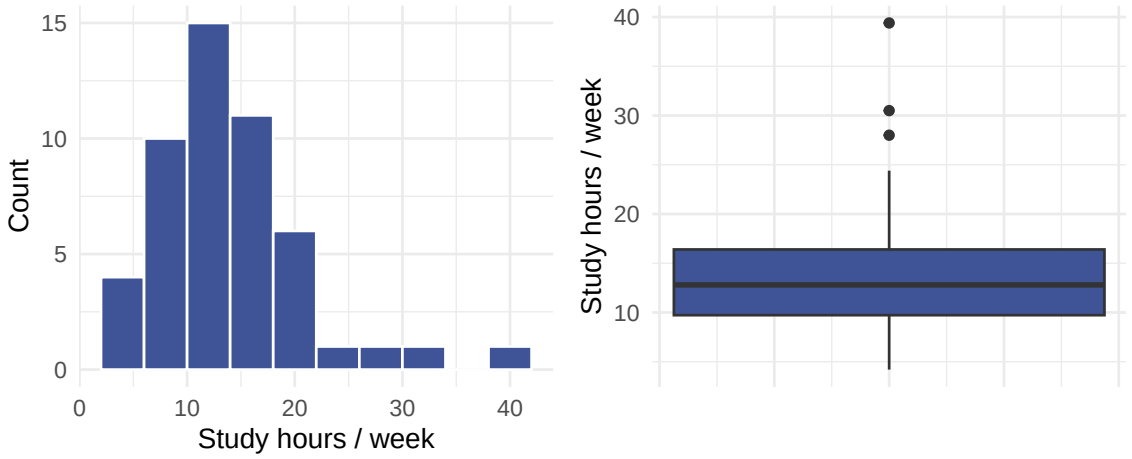


Figure 5.7: Weekly study hours: histogram (left) and box plot (right). The distribution is right-skewed with one high outlier.

---

4. Compare a numerical variable across groups

To see whether a numerical variable differs between groups, use **side-by-side box plots** and a **grouped summary**. We'll compare `commute_min` for On-campus vs. Off-campus students.

5.5.10 StatLens

1. Open the Grouped Statistics explorer.
2. Set the response to **`commute_min`** and the grouping variable to **`housing`**. Compare the two box plots and the group medians.

5.5.11 Jamovi

1. **Exploration → Descriptives.** Put `commute_min` in **Variables** and `housing` in **Split by**.
2. Request **Median** and **IQR**; under **Plots**, turn on **Box plot**.

(*Jamovi screenshots to be added.*)

5.5.12 R

```
survey |>
  group_by(housing) |>
  summarize(median = median(commute_min), IQR = IQR(commute_min), n = n())
```

```
# A tibble: 2 x 4
  housing    median   IQR     n
  <chr>       <dbl> <dbl> <int>
1 Off-campus     16  23.2    28
2 On-campus       6   3      22
```

```
ggplot(survey, aes(x = housing, y = commute_min, fill = housing)) +
  geom_boxplot(show.legend = FALSE) +
  labs(x = "Housing", y = "Commute (min, one-way)") +
  theme_minimal(base_size = 12)
```

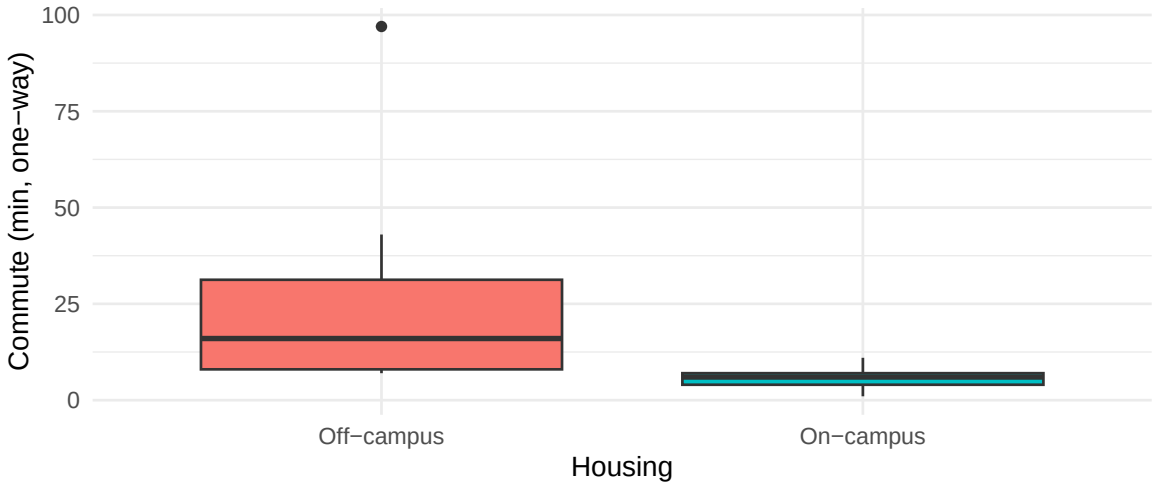


Figure 5.8: One-way commute time for on-campus vs. off-campus students.

---

5. Explore a relationship between two numerical variables

For two numerical variables, use a **scatterplot** and the **correlation** `r`. We'll look at `study_hours` vs. `sleep_hours`. Not every pair is related — part of the skill is reading a weak relationship honestly.

5.5.13 StatLens

1. Open the Regression / Scatterplot explorer.
2. Read the **scatterplot** and the reported **correlation r**. Is there a clear trend, or only a weak one?

5.5.14 Jamovi

1. **Regression → Correlation Matrix.** Add `study_hours` and `sleep_hours`.
2. Read **Pearson's r**. For the plot, use **Exploration → Scatterplot**.

(*Jamovi screenshots to be added.*)

5.5.15 R

```
# Correlation coefficient
survey |> summarize(r = round(cor(study_hours, sleep_hours), 2))
```

```
# A tibble: 1 x 1
      r
  <dbl>
1 -0.05
```

```r
ggplot(survey, aes(x = study_hours, y = sleep_hours)) +
  geom_point(color = "#3E5496", size = 2) +
  labs(x = "Study hours / week", y = "Sleep (hours / night)") +
  theme_minimal(base_size = 12)
```

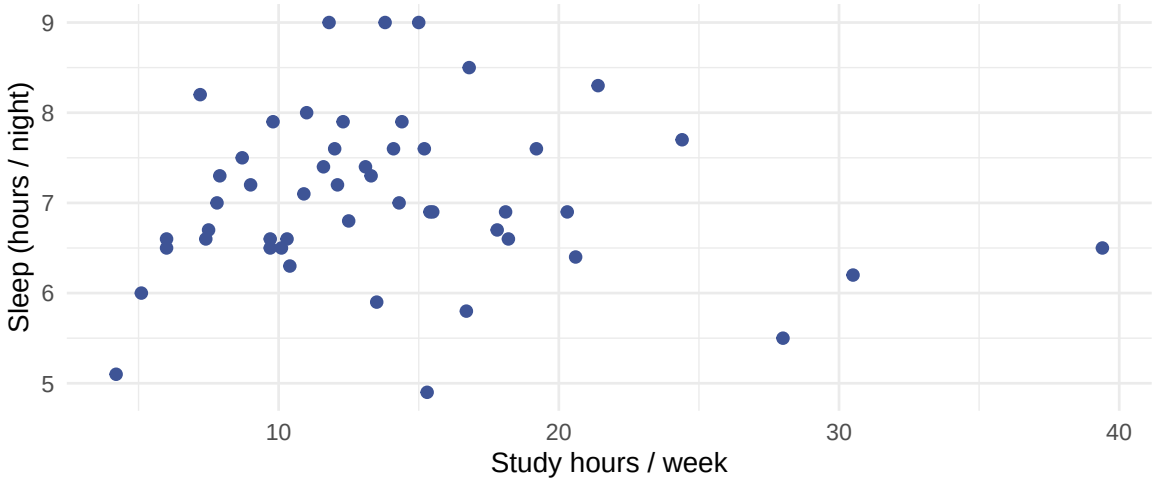


Figure 5.9: Study hours vs. nightly sleep. The cloud of points shows little association.

Going further in R (optional)

If you chose the R path and want guided, hands-on practice, the OpenIntro *IMS* interactive tutorials cover this unit in depth (free, runs in your browser):

- Language of data
- Visualizing categorical data
- Visualizing numerical data
- Summarizing with statistics
- Visualizing two variables

They use the same `tidyverse` style as the R tab above.

Check yourself

You can now, in your chosen tool:

- ☐ open a data file and classify each variable as categorical or numerical;
- ☐ make a frequency table and bar chart for a categorical variable;
- ☐ make a histogram and box plot and report center and spread for a numerical variable;
- ☐ compare a numerical variable across groups with side-by-side box plots;
- ☐ make a scatterplot and report the correlation for two numerical variables.

These are exactly the moves you'll need for the **Explore project** that follows.

Project: Exploring a Distribution

Project — Exploratory Data Analysis. A *project* is a short analysis you carry out end to end and **communicate** as a brief report. Do the work in **Jamovi** (or **R**) and submit the deliverable described at the end. The skill here is producing the *right* summaries and reading them honestly.

The question

Collectors buy and sell the game *Mario Kart* (Nintendo Wii) in online auctions. You have the final prices from a sample of completed auctions.

> **What is a typical selling price, how much do prices vary, and do *new* and *used* copies sell differently?**

The data

The `mariokart` dataset holds 143 eBay auctions. You'll use `total_pr` (final price, in dollars) and `cond` (`new` or `used`). A couple of auctions were bundles sold at unusually high prices — keep an eye out for **outliers**.

Plan it first

a. **Variable types.** Classify `total_pr` and `cond`.
b. **Displays & summaries.** To describe *one numerical* variable, which display and which numerical summaries are appropriate?
c. **Resistance.** Given the possible high-price outliers, which measures of center and spread will you *trust* — mean/SD or median/IQR — and why?
d. **Comparison.** To compare price for `new` vs. `used`, which display fits a numerical variable split by a categorical one?

Do the analysis

5.5.16 Jamovi

1. Open `mariokart.csv` (**Open → Data**).
2. **Exploration → Descriptives.** Put `total_pr` in **Variables** and `cond` in **Split by**. Request **mean, median, std. deviation, IQR**.
3. Under **Plots**, turn on a **histogram** and a **box plot**.
4. Read off the typical price, the spread, and how new vs. used compare.

(*Jamovi screenshots to be added.*)

5.5.17 R

```r
library(dplyr)
library(ggplot2)
library(patchwork)
library(openintro)
data(mariokart)

# Overall numerical summary (note how outliers pull the mean above the median)
mariokart |>
  summarize(mean = mean(total_pr), median = median(total_pr),
            sd = sd(total_pr), IQR = IQR(total_pr))
```

```
# A tibble: 1 x 4
   mean median    sd   IQR
  <dbl>  <dbl> <dbl> <dbl>
1  49.9   46.5  25.7  12.8
```

```r
# Center & spread by condition (median/IQR are resistant to the outliers)
mariokart |>
  group_by(cond) |>
  summarize(median = median(total_pr), IQR = IQR(total_pr), n = n())
```

```
# A tibble: 2 x 4
  cond  median   IQR     n
  <fct>  <dbl> <dbl> <int>
1 new     54.0  9.75    59
2 used    42.8  8.91    84
```

```r
p1 <- ggplot(mariokart, aes(total_pr)) +
  geom_histogram(binwidth = 15, fill = "#3E5496", color = "white") +
  labs(x = "Final price ($)", y = "Count", title = "All auctions") +
  theme_minimal(base_size = 12)

p2 <- ggplot(mariokart, aes(cond, total_pr, fill = cond)) +
  geom_boxplot(show.legend = FALSE) +
  labs(x = "Condition", y = "Final price ($)", title = "By condition") +
  theme_minimal(base_size = 12)

p1 + p2
```

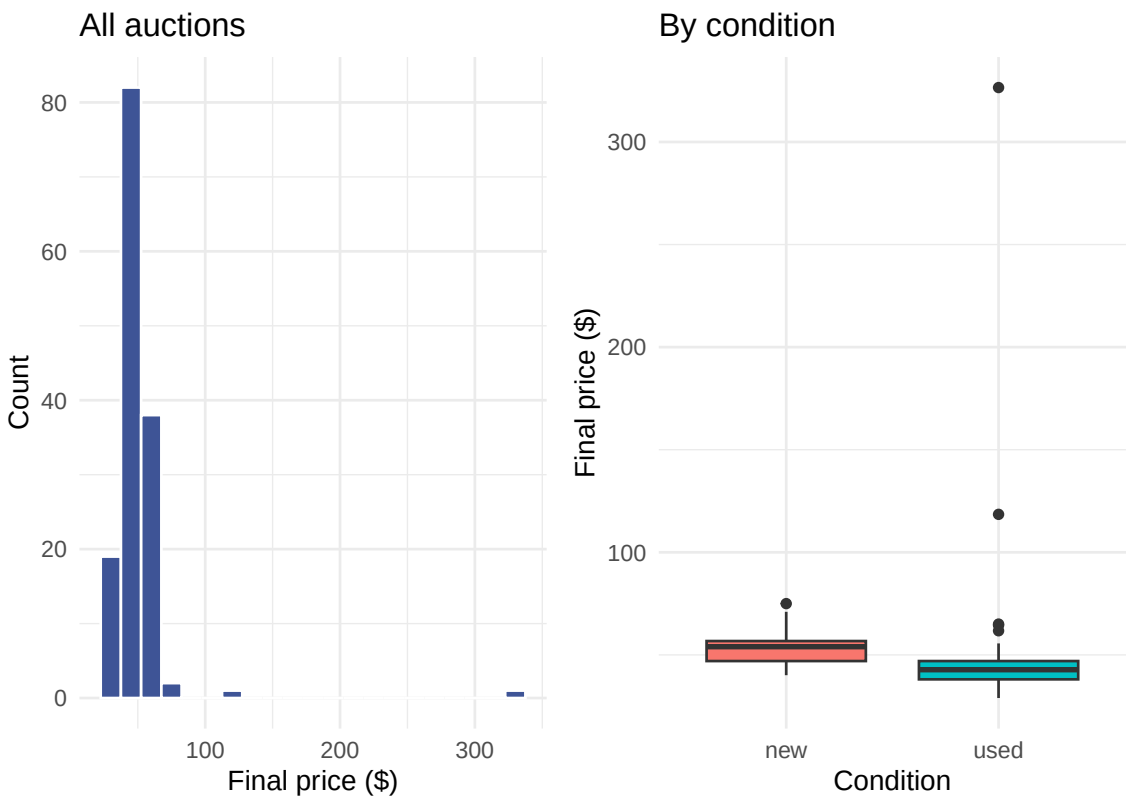


Figure 5.10: Distribution of Mario Kart auction prices (left) and prices split by condition (right). A couple of high-priced bundle auctions stretch the right tail.

Your deliverable

Submit a short report containing **all** of:

a. **Graphic** — a labeled histogram of `total_pr` (and, for the comparison, side-by-side boxplots by `cond`).
b. **Center & spread** — report the median and IQR, and say why you chose them over the mean and SD here.
c. **Shape & outliers** — describe the shape (symmetric? skewed?) and identify any outliers, with a sentence on what they are.
d. **Comparison in context** — 2–3 sentences comparing `new` vs. `used` prices.
e. **One limitation** — e.g., these are only completed auctions of one game on one platform.

How it's graded

Criterion	What we look for
Appropriate, labeled graphic	histogram + boxplots, axes labeled
Right summaries for the data	median/IQR chosen given the outliers, justified
Shape & outliers described	correct shape; outliers identified
Comparison **in context**	new vs. used stated in plain language
Communication	clear, correct, no over-claiming

Warm up first. Explore the same kind of summaries interactively in the Descriptive Statistics explorer before building your report.

PART II

UNIT 2: SIMULATE

How can we draw conclusions about a population from a single sample? We answer this question through *simulation*. By resampling data thousands of times, we build an intuitive understanding of sampling variability, hypothesis testing, and confidence intervals — without formulas or distribution tables.

This is the heart of the simulation-first approach: understand the *logic* of inference before learning the mathematical shortcuts.

Chapter 6

Sampling Variability

So far in this course, you have learned how to collect data, visualize it, and compute sample statistics. But statistics is about more than describing the data you have — it is about using data to learn about a larger population. The central challenge is this: **every sample is different**. If you took another sample from the same population, you would get different results. This chapter introduces the foundational ideas that make statistical inference possible: the distinction between parameters and statistics, the concept of a sampling distribution, and the standard error as a measure of how much sample statistics vary. These ideas are the bridge from describing data to drawing conclusions about the world.

Two quick warmups: how good is your intuition about randomness?

- **Are You Psychic?** — predict a series of coin flips, see your score, then watch the tool simulate hundreds of random guessers. How impressive is *your* result compared to pure chance?
- **Stump the Chump** — type an H/T sequence you think "looks random," then compare it to a truly random sequence. Run lengths, alternation rates, and the longest run will tell you whether human-generated "randomness" is too even.

Both are short interactive demos that motivate the rest of this chapter: real random samples vary more than people expect, and that variability has a predictable shape.

6.1 Parameters vs. statistics

When we collect data, we typically have a larger group in mind. A political poll surveys 1,000 likely voters, but the real interest is in all voters. A quality engineer measures 50 widgets from an assembly line, but the real interest is in all widgets the line produces. The larger group is the **population**; the subset we actually observe is the **sample**.

A **population parameter** is a numerical summary of the entire population. It is a fixed number, but it is almost always unknown because we cannot observe the whole population.

A **sample statistic** (also called a **point estimate**) is a numerical summary computed from sample data. It is our best guess at the population parameter. Because the statistic is computed from a sample, it is known — but it changes from sample to sample.

We use specific notation to distinguish parameters from statistics:

Quantity	Parameter (population)	Statistic (sample)
Proportion	p	$\hat{p}$

Quantity	Parameter (population)	Statistic (sample)
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Difference in proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
Difference in means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$

The "hat" notation ($\hat{p}$) is a common convention: it signals that we are estimating the corresponding parameter.

A medical consultant has facilitated 62 liver donor surgeries, and 3 of her patients experienced complications. She uses this track record to attract new patients.

Let p represent the true complication rate for liver donors working with this consultant. What is the point estimate $\hat{p}$?

The sample proportion is $\hat{p} = 3/62 = 0.048$. That is, about 4.8% of her patients experienced complications. This is our point estimate of the true complication rate p.

A Stanford University graduate student recruited 120 tappers and 120 listeners for a study on communication. Only 3 out of 120 listeners were able to guess the tune being tapped. What is the point estimate $\hat{p}$ for the true proportion of listeners who can guess correctly?

Show answer

The point estimate is $\hat{p} = 3/120 = 0.025$, or 2.5%. This is our best single estimate of the true proportion p of listeners who can guess the tune.

In a study of sex discrimination, 48 male bank supervisors each reviewed a personnel file and decided whether the candidate should be promoted. The files were identical except that half listed the candidate as male and half as female. The results are summarized below:

	Promoted	Not promoted	Total
Male	21	3	24
Female	14	10	24
Total	35	13	48

What is the point estimate for the difference in promotion rates between male and female candidates?

The promotion rate for male candidates is $\hat{p}_M = 21/24 = 0.875$ and for female candidates is $\hat{p}_F = 14/24 = 0.583$. The point estimate of the difference is:

$$\hat{p}_M - \hat{p}_F = 0.875 - 0.583 = 0.292$$

The observed difference is 29.2 percentage points in favor of male candidates.

A point estimate is almost never exactly equal to the population parameter. This is not because we did anything wrong — it is an inherent consequence of working with a subset of the population. The critical question is: *how far off might the estimate be?*

6.2 Sampling distributions

To understand how far a statistic might be from its parameter, we need to understand what happens when we take **many samples** from the same population.

Imagine a large bowl containing thousands of beads, 60% red and 40% white. The proportion of red beads in the bowl is the population parameter: $p = 0.60$. Now suppose you reach in and scoop out a sample of $n = 50$ beads. You count 28 red beads, giving $\hat{p} = 28/50 = 0.56$. Your friend scoops a different sample of 50 and counts 32 red, giving $\hat{p} = 32/50 = 0.64$. Another friend gets 30 red for $\hat{p} = 0.60$.

Each person got a different sample proportion. No one did anything wrong — this is **sampling variability** in action.

Now imagine hundreds of people each scooping their own sample of 50 beads. If we recorded every $\hat{p}$ and made a histogram of all those values, we would see the **sampling distribution** of $\hat{p}$.

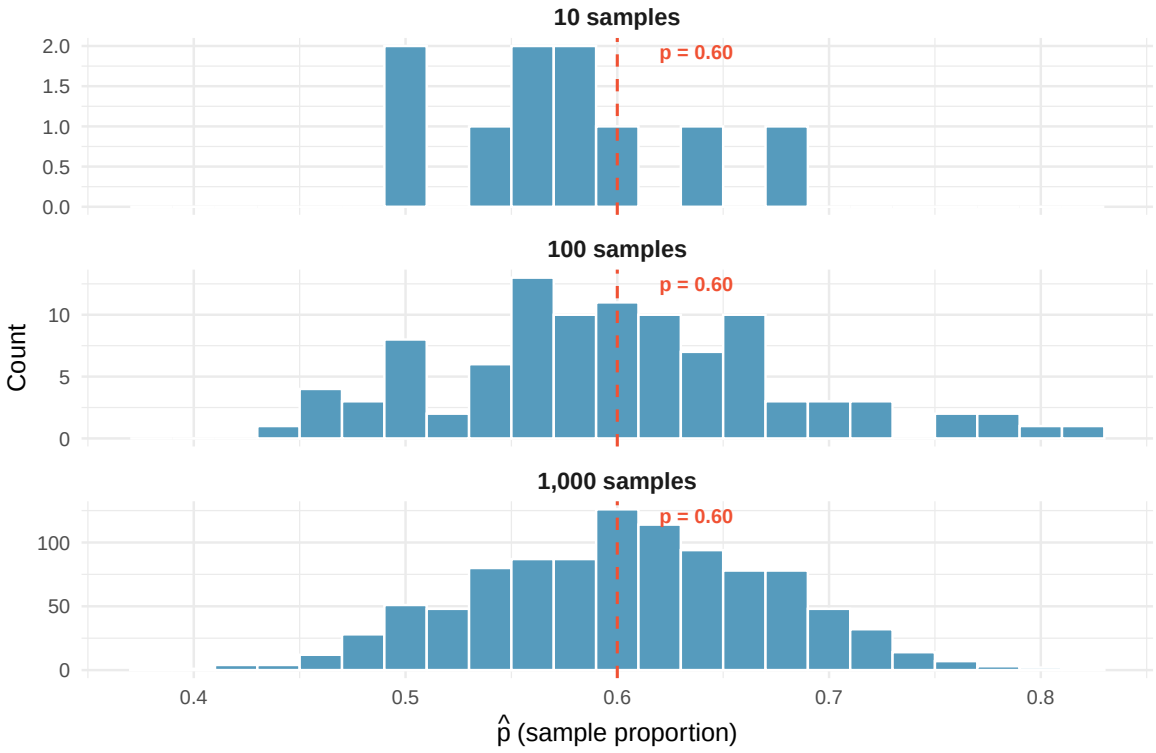


Figure 6.1: Building a sampling distribution. Each dot represents the proportion of red beads from one sample of n = 50. With more samples, the bell-shaped pattern centered at p = 0.60 becomes clear. ↗ Try this live — draw your own samples in the Sampling Distribution Lab and watch the picture fill in.

The **sampling distribution** of a statistic is the distribution of all possible values of that statistic computed from all possible samples of a given size from a given population.

A sampling distribution is *not* the distribution of individual data values. It is the distribution of a *sample statistic* (like $\hat{p}$ or $\bar{x}$) across many hypothetical samples.

Key properties of sampling distributions. For many common statistics, the sampling distribution has three important features:

1. **Center**: The sampling distribution is centered at the population parameter. On average, the sample statistic equals the parameter.
2. **Spread**: The sampling distribution has a measurable spread that decreases as sample size increases. Larger samples give more precise estimates.
3. **Shape**: For many statistics, the sampling distribution is approximately bell-shaped (normal) when the sample size is large enough.

Here is an analogy that may help. You want to know the average depth of a lake.

- **One boat, one sample.** Send a single boat out to measure the depth at 10 random locations on the lake. The captain reports back the 10 individual depths. Plotting those 10 numbers gives you a snapshot of the *data distribution* — how depth varies from spot to spot on this lake. If you sent that same boat out again to 10 new random spots, you would get 10 different numbers, but the shape of the picture would look similar. The **average** of the 10 measurements is one estimate of the true mean depth — a single sample statistic.
- **A fleet of boats, a sampling distribution.** Now send a fleet of 100 boats out, each to its own 10 random locations. Each captain reports back only the **average** of their 10 depths — one number per boat. Plotting the 100 averages gives you the *sampling distribution of the mean*. It is centered at the true lake depth, but it is much narrower than the data distribution, because averaging pulls in the extremes.

The data distribution and the sampling distribution describe fundamentally different things: how single depths vary across the lake versus how sample means vary across repeated samples of size 10. Notice too that a single boat that stays in one bay all day is *not* a random sample from the lake — the 10 depths it reports would be a **cluster sample**, and averaging them would give a systematically biased picture of the whole lake. A proper sampling distribution comes from imagining many *random* samples, not many *convenient* ones.

The two distributions are fundamentally different:

- The **data distribution** shows how individual observations vary.
- The **sampling distribution** shows how a *statistic* (like the mean) varies from sample to sample.

6.3 Variability of sample statistics

Different samples from the same population produce different statistics. This variability is not a flaw — it is a fundamental feature of sampling that we must understand and quantify.

Suppose the true proportion of all US adults who support a particular policy is $p = 0.52$. Three polling organizations each take a random sample of $n = 1{,}000$ adults and compute $\hat{p}$. Their results might look like:

Poll	Sample proportion $\hat{p}$
Poll A	0.49
Poll B	0.54
Poll C	0.51

All three polls were conducted correctly, yet they produced different results. Poll A even shows less than 50% support! This is sampling variability: the unavoidable variation that occurs when we estimate a population parameter from a sample.

If the three polls above each used a sample of $n = 10{,}000$ instead of $n = 1{,}000$, would you expect the three $\hat{p}$ values to be closer together or farther apart? Why?

Show answer

We would expect them to be closer together. Larger samples contain more information about the population, so the statistic is more precise — meaning less variability from sample to sample.

6.3.1 What determines how much a statistic varies?

Two factors control how much a sample statistic bounces around from sample to sample:

1. **Sample size** (n): Larger samples produce statistics that are closer to the parameter. This is intuitive — a sample of 1,000 tells you more about the population than a sample of 10.
2. **Population variability**: If the population is very homogeneous (everyone is similar), then any sample will look like any other sample, and the statistic won't vary much. If the population is very heterogeneous (lots of individual differences), then different samples can look quite different, and the statistic will vary more.

Of these two, sample size is the one we can control, and it turns out to be very powerful. Quadrupling the sample size cuts the variability of most statistics roughly in half.

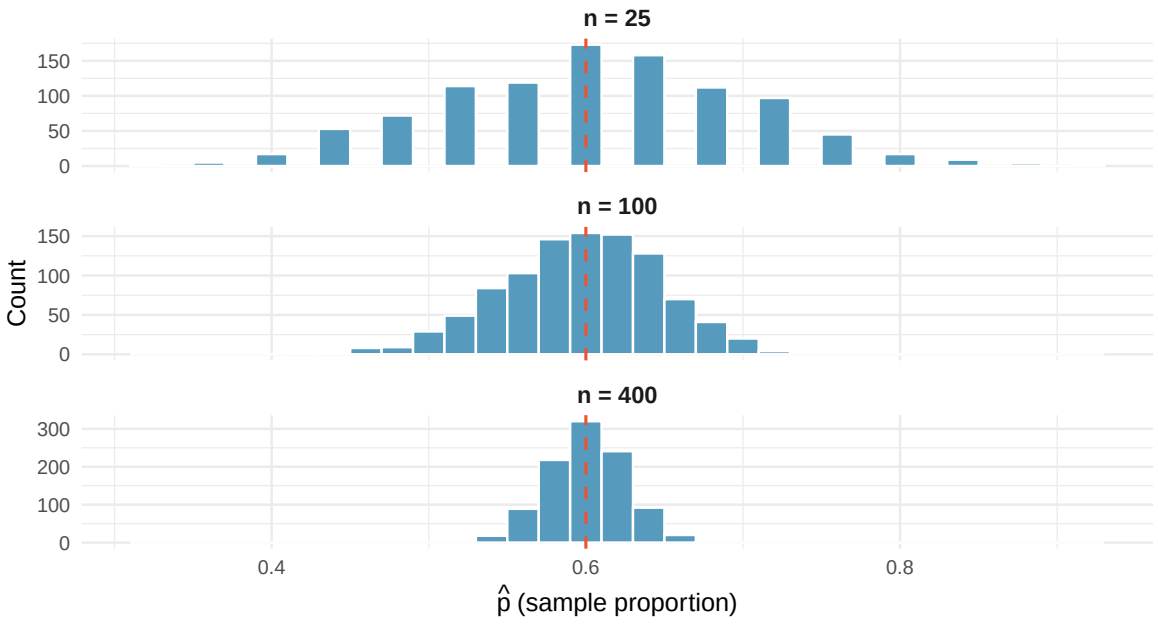


Figure 6.2: Sampling distributions of the sample proportion for three different sample sizes. All three are centered at p = 0.60, but the spread decreases dramatically as the sample size increases. ↗ Try this live — drive the n slider yourself in the Sampling Distribution Lab and watch the spread shrink as n grows.

Figure 6.2 shows this dramatically: with $n = 25$, individual sample proportions range from about 0.35 to 0.85. With $n = 400$, they are tightly clustered between 0.55 and 0.65.

Predict → Do → Explain. The three properties above — center, spread, and shape — hold for *any* statistic. Open the Sampling Distribution Lab to build the sampling distribution of the **sample mean** from a right-skewed population. The gated activity opens with a **Predict** step (commit to where the sample means will pile up *before* you draw), walks you through drawing samples of $n = 10$ — a few one at a time, then a thousand — and then freezing that picture to compare it with $n = 100$, with a reality-check gate at each stage, and closes with an **Explain** step where you articulate why the sampling distribution ends up centered at μ and roughly bell-shaped even when the population is

not. Two distinctions the lab is designed to make crisp: **more samples fills in the same distribution**, but **larger n produces a different, narrower one**. Your predictions and explanation save locally in the tool, so you can return to your work on the same device.

> **i** Self-check
>
> At $n = 10$ the sampling distribution is already less skewed than the population but still somewhat right-skewed and fairly wide; by $n = 100$ it is much narrower and clearly bell-shaped — a first glimpse of the **Central Limit Theorem**: averages of independent draws become approximately normal as n grows, even when the population is not. The center stays put at the population mean because averaging is unbiased — no amount of skew shifts it. The spread shrinks because $\sigma/\sqrt{n}$ shrinks: quadrupling n cuts the standard error roughly in half.

6.4 Standard error

We need a single number that quantifies how much a statistic varies from sample to sample. That number is the **standard error**.

The **standard error** (SE) of a statistic is the standard deviation of its sampling distribution. It measures the typical distance between a sample statistic and the population parameter.

A small standard error means the statistic tends to fall close to the parameter (precise estimate). A large standard error means the statistic could plausibly fall far from the parameter (imprecise estimate).

Do not confuse the standard error with the standard deviation of the data:

- The **standard deviation** (s or σ) measures how much *individual observations* vary around the mean.
- The **standard error** (SE) measures how much a *sample statistic* varies around the population parameter from sample to sample.

The standard error is always smaller than the standard deviation (for samples larger than $n = 1$) because averaging reduces variability.

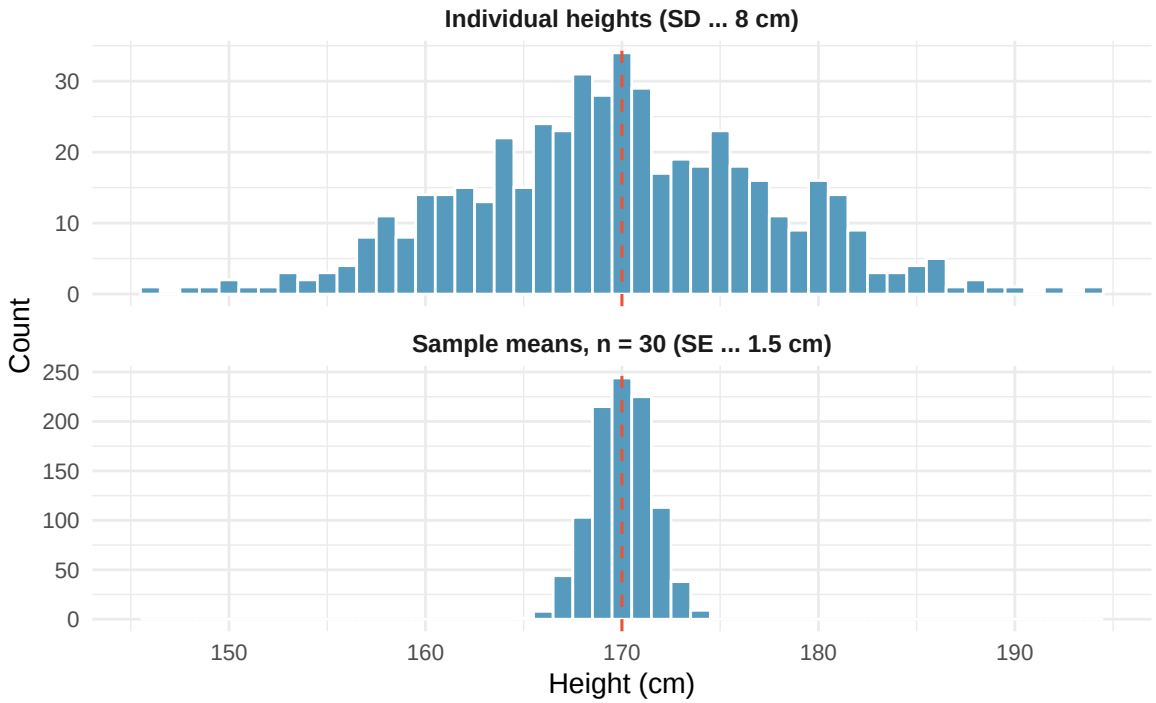


Figure 6.3: Comparing the data distribution (top) with the sampling distribution of the mean (bottom). Both are centered near the same value, but the sampling distribution is much narrower — its spread is the standard error, not the standard deviation. ↗ Try this live — build the same population-versus-sample-means comparison from a bell-shaped population in the Sampling Distribution Lab, and vary n to watch the sampling distribution narrow further.

In Figure 6.3, the top panel shows how individual heights vary (SD ≈ 8 cm), while the bottom panel shows how sample means vary (SE ≈ 1.5 cm). Both are centered near $\mu = 170$, but the sampling distribution of the mean is much narrower.

In the medical consultant example, $\hat{p} = 3/62 = 0.048$. If we could take many samples of 62 patients from this consultant's true patient population and compute $\hat{p}$ each time, the standard deviation of all those $\hat{p}$ values would be the standard error of $\hat{p}$.

While we usually cannot literally take repeated samples, simulation methods (which we will explore in the randomization tests and bootstrap confidence intervals chapters) allow us to approximate the standard error. In this case, bootstrap simulations suggest that the standard error of $\hat{p}$ is approximately 0.027. This means that if we could repeat the study many times, the sample complication rate would typically differ from the true complication rate by about 2.7 percentage points.

6.4.1 Why the standard error matters

The standard error is the engine of statistical inference. Here is why:

- **Confidence intervals** are built by going a certain number of standard errors above and below the point estimate. A rough 95% confidence interval is:

$$\text{statistic} \pm 2 \times SE$$

- **Hypothesis tests** compare an observed statistic to what we would expect under the null hypothesis. The comparison is made in units of standard errors: "Is the observed result more than 2 standard errors from the null value?"
- **Sample size planning** uses the standard error to determine how many observations are needed to achieve a desired level of precision.

In the coming chapters, we will use simulation to *estimate* the standard error. In later chapters (Unit 3), we will use mathematical formulas to *calculate* it directly.

Two researchers both estimate the proportion of college students who prefer online textbooks. Researcher A surveys 100 students and finds $\hat{p}_A = 0.62$ with $SE_A = 0.049$. Researcher B surveys 400 students and finds $\hat{p}_B = 0.58$ with $SE_B = 0.025$.

(a) Whose estimate is more precise? How do you know?
(b) Why is Researcher B's standard error smaller?

Show answer

(a) Researcher B's estimate is more precise because the standard error is smaller ($0.025 < 0.049$), meaning the sample statistic is expected to be closer to the true parameter. (b) Researcher B surveyed four times as many students, and increasing the sample size reduces the standard error.

6.5 Looking ahead

This chapter has established the core vocabulary and ideas that drive all of statistical inference:

- We want to know about **parameters** but can only observe **statistics**.
- Statistics vary from sample to sample — this is **sampling variability**.
- The pattern of this variation is described by the **sampling distribution**.
- The **standard error** quantifies the typical amount of sampling variability.

The fundamental question of inference is: *Given the sampling variability we expect, is the pattern in our data real, or could it have happened by chance?*

In the next three chapters, we will answer this question using three different approaches — all rooted in simulation:

- **Randomization Tests**: Shuffle the data to see what differences look like when there is no real effect. If the observed difference is extreme compared to the shuffled results, we have evidence of a real effect.
- **Bootstrap Confidence Intervals**: Resample from the data to approximate the sampling distribution, and use it to construct a range of plausible values for the parameter.
- **Sampling Distributions via Simulation**: Connect the simulation results from the randomization tests and bootstrap confidence intervals chapters to mathematical theory, setting the stage for the formula-based methods in Unit 3.

The goal of Unit 2 is to build your *intuition* for inference. The formulas will come later; the ideas come first.

6.6 Chapter review

6.6.1 Summary

- A **population parameter** is a fixed but unknown numerical summary of the population. A **sample statistic** (point estimate) is computed from data and serves as our best guess of the parameter.
- Different samples from the same population yield different statistics. This is **sampling variability** — it is not a mistake, but a fundamental feature of working with samples.
- The **sampling distribution** of a statistic describes how the statistic varies across all possible samples of a given size. It is typically centered at the parameter, and its spread decreases as sample size increases.
- The **standard error** is the standard deviation of the sampling distribution. It measures the precision of the statistic as an estimate of the parameter.
- The standard error is not the same as the standard deviation of the data. The standard deviation measures variability of individual observations; the standard error measures variability of a statistic.
- Larger samples produce smaller standard errors, meaning more precise estimates.

6.6.2 Key terms

Term	Definition
Population parameter	A fixed numerical summary of the entire population (e.g., p, μ)
Sample statistic	A numerical summary computed from sample data (e.g., $\widehat{p}$, $\bar{x}$)
Point estimate	Another name for sample statistic; our best guess at the parameter
Sampling variability	The fact that different samples produce different statistics
Sampling distribution	The distribution of a statistic over all possible samples of a given size
Standard error (SE)	The standard deviation of the sampling distribution

6.7 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Identify the parameter, I.** For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

a. In a survey, 100 college students are asked how many hours per week they spend on the Internet.

b. In a survey, 100 college students are asked: "What percentage of the time you spend on the Internet is part of your course work?"

c. In a survey, 100 college students are asked whether they cited information from Wikipedia in their papers.

d. In a survey, 100 college students are asked what percentage of their total weekly spending is on alcoholic beverages.

e. In a sample of 100 recent college graduates, it is found that 85 percent expect to get a job within one year of their graduation date.

2. **Identify the parameter, II.** For each of the following situations, state whether the parameter of interest is a mean or a proportion.

a. A poll shows that 64% of Americans personally worry a great deal about federal spending and the budget deficit.

b. A survey reports that local TV news has shown a 17% increase in revenue within a two year period while newspaper revenues decreased by 6.4% during this time period.

c. In a survey, high school and college students are asked whether they use geolocation services on their smart phones.

d. In a survey, smart phone users are asked whether they use a web-based taxi service.

e. In a survey, smart phone users are asked how many times they used a web-based taxi service over the last year.

3. **Sample size and the sampling distribution.** A researcher plans to estimate the mean sleep hours per night for full-time college students in the U.S. She considers three possible sample sizes: $n = 25$, $n = 100$, and $n = 400$. In each case, she would draw a simple random sample and compute $\bar{x}$; the sampling distribution describes how $\bar{x}$ varies across all possible samples of that size.

a. As n grows from 25 to 400, what happens to the **center** of the sampling distribution? Why?

b. As n grows from 25 to 400, what happens to the **spread** (standard error) of the sampling distribution? Give a rough factor: quadrupling n from 25 to 100 changes the standard error by roughly what?

c. As n grows from 25 to 400, what happens to the **shape** of the sampling distribution (assuming the population is not extremely skewed)? Describe what you would expect to see if you drew the histogram at each sample size.

d. Which of the three sample sizes gives the most precise estimate of the population mean? Which requires the most resources to collect? Briefly discuss the trade-off.

4. **Statistic vs. parameter with correct notation.** For each described study, identify the population parameter and the sample statistic. For each, state whether it is a **mean** or a **proportion**, and write it with the correct symbol.

a. A national poll of 800 US adults finds that 42% support a proposed carbon tax. The pollster wants to estimate the true percentage of all US adults who support the tax.

b. A researcher measures the resting heart rate of 60 randomly-selected adult women in a city and finds a sample mean of 71 beats per minute. She wants to estimate the mean resting heart rate for the population of all adult women in the city.

c. A company tests 200 randomly-selected components off the assembly line and finds 5 defective. The quality engineer wants to estimate the long-run defect rate on this line.

d. In each of (a)–(c), which of the two — parameter or statistic — is known exactly, and which is being estimated?

5. **Data distribution or sampling distribution?** For each of the following descriptions, decide whether the histogram would show a **data distribution** (the distribution of individual observations) or a **sampling distribution** (the distribution of a sample statistic across many samples). Justify each in one sentence.

a. A histogram of the ages of all 1,200 patients admitted to a hospital ward last year.

b. A histogram of the mean age from each of 500 random samples of 30 patients drawn from those 1,200.

c. A dotplot showing the proportion of red beads for each of 400 people, where each person independently scooped a sample of 50 beads from the same bowl.

d. A histogram of the SAT scores from a random sample of 250 high-school seniors.

e. Which of the four would typically be **narrower** — a data distribution or a sampling distribution built from the *same* population? Explain in one sentence.

6. **Reading a sampling distribution.** The histogram below was built by drawing 1,000 random samples of size $n = 40$ from a large population and recording the sample proportion $\hat{p}$ (the proportion of "yes" responses to a survey question) for each. The true population proportion is $p = 0.50$.

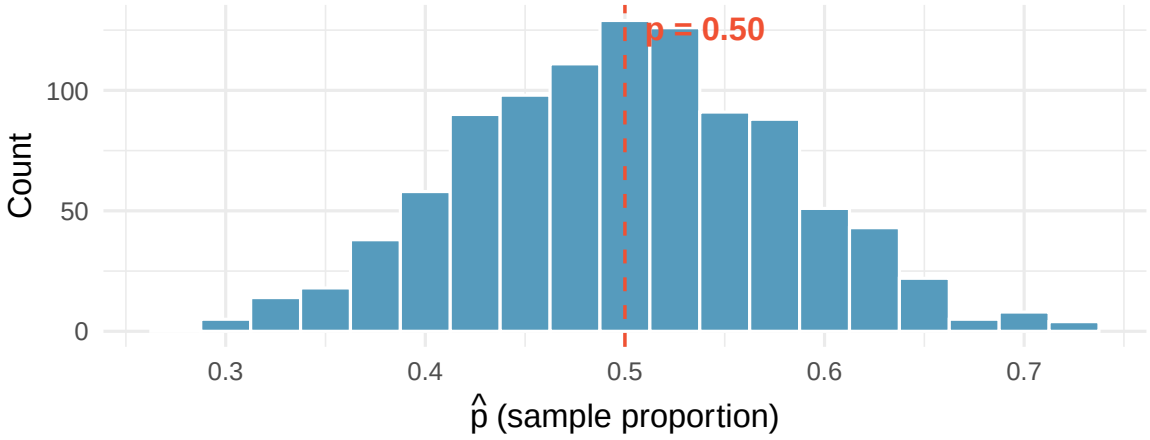


Figure 6.4: 1,000 sample proportions $\hat{p}$ from samples of size $n = 40$, drawn from a population with $p = 0.50$.

a. Where is the sampling distribution **centered**? How does that compare to the true p?

b. Describe the **shape** of the sampling distribution.

c. Estimate the **standard error** of $\hat{p}$ by eyeballing the picture. (Recall: the SE is the *standard deviation* of the sampling distribution — roughly, the typical distance from the center.)

d. Suppose the samples were instead of size $n = 160$ (four times as large). Describe how the picture would change. In particular, roughly what would the new standard error be?

↗ Try this yourself — the Sampling Distribution Lab lets you regenerate this picture live and drag the sliders to see what happens at other n or p.

7. **Standard deviation vs. standard error.** A study of 400 college students reports sample mean sleep $\bar{x} = 6.8$ hours per night, sample standard deviation $s = 1.4$ hours, and standard error of the mean SE$(\bar{x}) = 0.07$ hours. For each of the following statements, decide whether s or SE$(\bar{x})$ is the appropriate number, and explain briefly.

a. "In this sample, a typical student's sleep is within a few hours of 6.8."

b. "If we drew a fresh sample of 400 college students and recomputed the mean, the new $\bar{x}$ would typically be within a tenth of an hour of the true population mean."

c. A researcher writes: "The variability in sleep hours across students in this sample is about 1.4 hours." Is she describing s or SE$(\bar{x})$?

d. Explain why SE$(\bar{x})$ is so much smaller than s for this study.

8. **What a reported standard error tells you.** A national polling firm surveys 1,067 US adults and reports that 51% support a particular ballot measure, with a margin of error of ± 3 percentage points at 95% confidence (which corresponds roughly to $2 \cdot \text{SE} \approx 0.03$, i.e., $\text{SE} \approx 0.015$).

a. If the firm's polling procedure were repeated many times with fresh random samples of 1,067 adults, roughly what would a picture of all their $\widehat{p}$ values look like? Describe the center and spread.

b. Would you be surprised if a different firm, using the same methodology and a fresh sample of 1,067 adults, reported 48% support? What about 40% support? Explain each briefly using the SE.

c. Suppose the firm doubled its sample size to about 2,134 adults. Roughly what would the new SE and margin of error become? (Hint: SE scales like $1/\sqrt{n}$, so doubling n multiplies SE by $1/\sqrt{2} \approx 0.71$.)

d. A political commentator sees the poll and writes: "Support is definitely between 48% and 54%." What's wrong with the word *definitely*?

9. **Match the sampling distributions.** Below are four histograms, each built by drawing 1,000 random samples and recording the sample proportion $\hat{p}$ each time. They come from four different (n, p) combinations:

- $(n = 25,\ p = 0.20)$
- $(n = 100,\ p = 0.20)$
- $(n = 25,\ p = 0.50)$
- $(n = 100,\ p = 0.50)$

Match each panel (A, B, C, D) to its combination, using only what you can see. Then briefly justify each match in one sentence: what feature of the picture told you *center* (relates to p) versus *spread* (relates to n)?

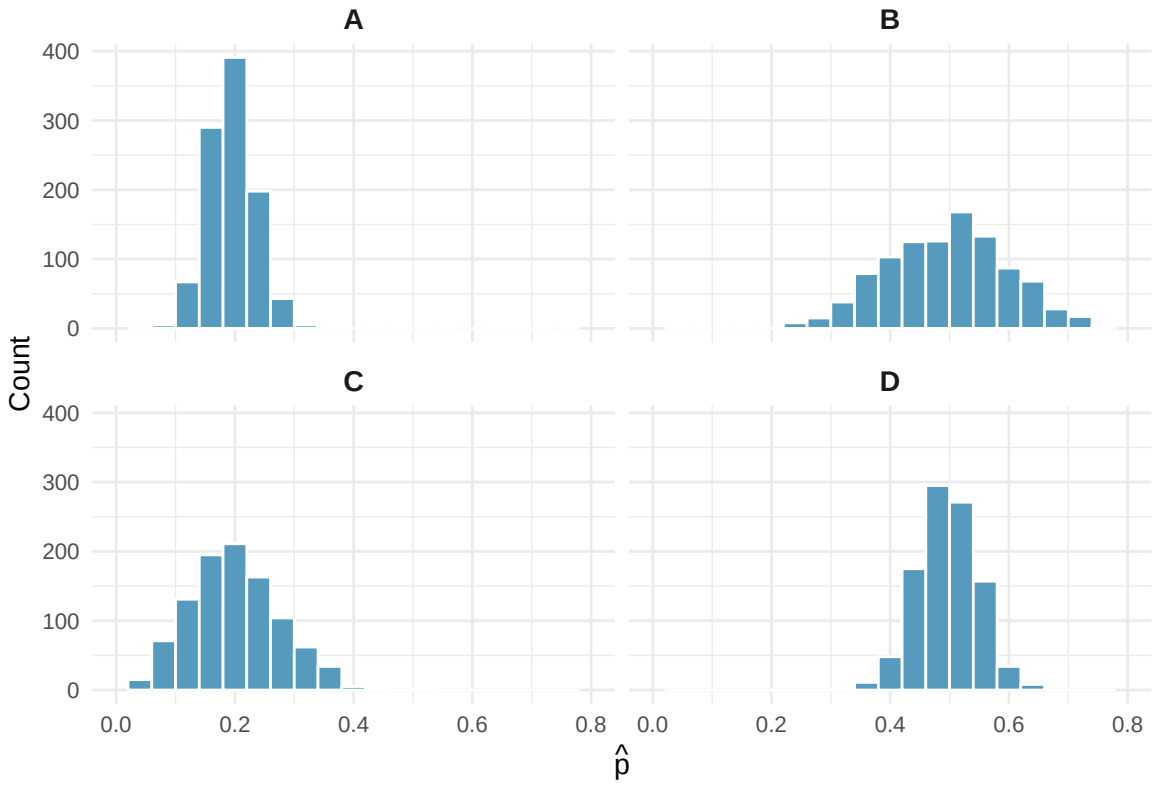


Figure 6.5: Four sampling distributions of $\hat{p}$. Each built from 1,000 random samples.

↗ Try this yourself — in the Sampling Distribution Lab, drive the n and p sliders to any of the four combos and watch the histogram redraw. Does it match the panel you assigned it to?

10. **Estimate the standard error from a dot plot.** The dot plot below shows the sample means $\bar{x}$ from 50 random samples of 30 college students each. In each sample, the researcher measured hours of sleep per night; the population mean is $\mu = 7.0$ hours.

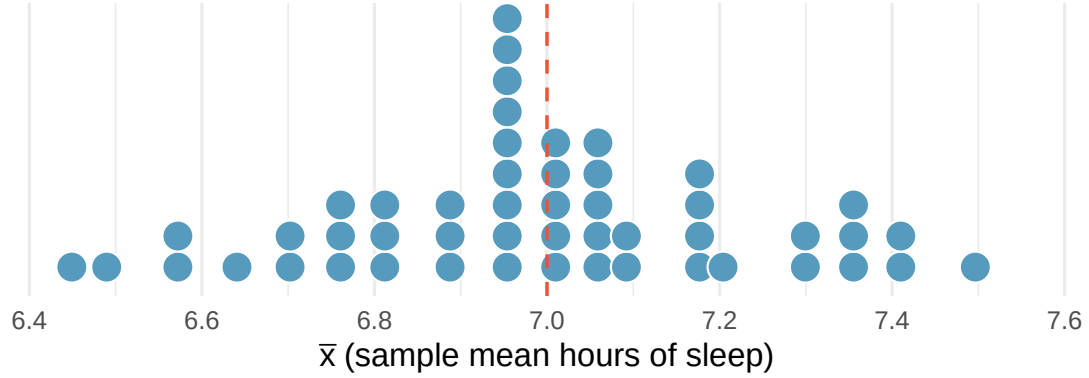


Figure 6.6: 50 sample means $\bar{x}$ from repeated samples of size $n = 30$. Population mean $\mu = 7.0$ hours.

a. Where is the sampling distribution of $\bar{x}$ **centered**? How does that compare to the population mean $\mu = 7.0$?

b. Recall: the standard error of $\bar{x}$ is the *standard deviation of the sampling distribution* — roughly, the typical distance of a dot from the center. Estimate the SE from the picture. (A rough eyeball: the range from lowest to highest dot spans about ± 2 SE if the shape is roughly bell-shaped.)

c. A student computes the standard error using the population SD $\sigma = 1.4$ hours: $\sigma/\sqrt{n} = 1.4/\sqrt{30} \approx 0.26$. How does that compare to your visual estimate in (b)?

d. If the researcher had drawn samples of size $n = 120$ instead of $n = 30$, how would the *width* of this dot plot change? Give a rough factor.

↗ Try this yourself — the Sampling Distribution Lab builds the same kind of picture from a bell-shaped population (its own μ and σ, not the sleep values); draw samples at $n = 30$, then drive n up to 120 and watch the spread shrink.

StatLens Exercises

These exercises focus on the ideas behind sampling variability: telling a **parameter** from a **statistic**, a **sampling distribution** from a **data distribution**, and reading the **standard error**. Where a link is given, let StatLens do the simulating and spend your effort on the reasoning.

1. **Parameter or statistic?** A university has 14,000 undergraduates. A researcher surveys a random sample of 200 of them and finds that 132 work at least 10 hours a week.

   a. Identify the population and the sample.
   b. Is the value $132/200 = 0.66$ a parameter or a statistic? Give its symbol.
   c. Is the true fraction of *all* 14,000 undergraduates who work 10+ hours a parameter or a statistic? Give its symbol. Is its value known?

2. **Sampling distribution vs. data distribution.** A nurse records the resting heart rate of every patient in a large clinic, then also computes the mean heart rate of each day's batch of 40 patients over many days.

   a. Which collection of numbers forms a *data distribution* — the individual heart rates, or the daily means?
   b. Which forms a *sampling distribution*?
   c. Which of the two distributions has the larger spread, and why?

3. **Predict, then check: two different knobs.** You saw the *effect of sample size* on the sampling distribution in the three-panel figure earlier in this chapter (all three centered at $p = 0.60$, spread shrinking as n grew). Now drive it yourself: *first, write your predictions; then* open the Sampling Distribution Simulator and choose the **right-skewed** population.

   a. **Predict** the shape of the sampling distribution of $\bar{x}$ for $n = 5$ (as skewed as the population, less skewed, or symmetric?). Then set $n = 5$, draw many samples, and compare to your prediction.
   b. With $n = 5$ still set, draw *even more* samples. Does the **spread** shrink, or does the distribution just fill in? Now switch to $n = 50$. Which of the two actions — *more samples* or *larger* n — changed the spread?
   c. The tool shows the theoretical standard error $\sigma/\sqrt{n}$ and the observed SD of the sample means. Why should these two numbers be close?

4. **Interpret the standard error.** A polling firm estimates support for a ballot measure from a random sample and reports $\hat{p} = 0.47$ with a standard error of 0.02.

   a. In plain language, what does the standard error of 0.02 tell you?
   b. Roughly how far from the true proportion might $\hat{p}$ reasonably be? (Use the "$\pm 2\,SE$" idea.)
   c. The firm wants to *halve* the standard error for its next poll. By what factor must it increase the sample size, and why?

5. **Spot the mistake.** A student writes: *"I took one sample of 60 students and got $\bar{x} = 7.1$ hours of sleep. The standard error was 0.3, so 95% of students sleep between about 6.5 and 7.7 hours."*

   a. What has the student confused?
   b. What does the interval $7.1 \pm 2(0.3)$ actually describe?

Chapter 7

Randomization Tests

Statistical inference is primarily concerned with understanding and quantifying the uncertainty of parameter estimates. While the equations and details change depending on the setting, the foundations for inference are the same throughout all of statistics. In this chapter, we introduce the **hypothesis testing framework** — a formal method for evaluating competing claims about a population using data. We learn how to build **randomization distributions** by shuffling data under the assumption that there is no effect, and we use these distributions to compute p**-values** that quantify the strength of the evidence.

7.1 The hypothesis testing framework

Throughout this course so far, you have worked with data in a variety of contexts. You have learned how to summarize and visualize data as well as how to describe relationships between variables. But more often than not, data have been collected to answer a research question about a larger group — and the data are a (hopefully) representative subset.

There is almost always variability in data — one dataset will not be identical to a second dataset even if they are both collected from the same population using the same methods. However, quantifying the variability in the data is neither obvious nor easy. Answering the question "*how* different is one dataset from another?" is not trivial.

Suppose your professor splits the students in your class into two groups: students who sit on the left side of the classroom and students who sit on the right side. If $\hat{p}_L$ represents the proportion of left-side students who prefer reading on a screen and $\hat{p}_R$ represents the proportion of right-side students who prefer reading on a screen, would you be surprised if $\hat{p}_L$ did not *exactly* equal $\hat{p}_R$?

While the proportions $\hat{p}_L$ and $\hat{p}_R$ would probably be close to each other, it would be unusual for them to be exactly the same. We would probably observe a small difference due to *chance*.

If we do not think the side of the room a person sits on in class is related to whether they prefer to read books on a screen, what assumption are we making about the relationship between these two variables?

Show answer

We would be assuming that these two variables are **independent** — knowing which side of the room someone sits on tells us nothing about their reading preference.

The question at the heart of hypothesis testing is: *Could the pattern we observe in the data have arisen from chance alone, or is there something real going on?*

Hypothesis testing.

A **hypothesis test** is a statistical technique used to evaluate competing claims using data. The process involves:

1. **State the hypotheses**: Define a null hypothesis (H_0) and an alternative hypothesis (H_A).
2. **Collect data**: Gather evidence that bears on the hypotheses.
3. **Assess the evidence**: Determine how likely the observed data would be if the null hypothesis were true.
4. **Draw a conclusion**: Decide whether the evidence is strong enough to reject the null hypothesis.

If the null hypothesis and the data notably disagree, then we reject the null hypothesis in favor of the alternative hypothesis. There are many nuances to hypothesis testing, and we will discuss these ideas throughout this chapter and those that follow.

7.2 Null and alternative hypotheses

The **null hypothesis** (H_0) often represents a skeptical perspective or a claim of "no difference" or "no effect." It assumes that any observed pattern in the data is due to chance alone.

The **alternative hypothesis** (H_A) represents the claim under investigation — typically that there *is* a real difference, a real effect, or a real relationship. The alternative hypothesis is usually the reason the research was conducted in the first place.

The US court system considers two possible claims about a defendant: they are either innocent or guilty. If we set these claims up in a hypothesis framework, which would be the null hypothesis and which the alternative?

The jury considers whether the evidence is so convincing that there is no reasonable doubt regarding the person's guilt. The skeptical perspective (null hypothesis) is that the person is innocent until evidence is presented that convinces the jury the person is guilty (alternative hypothesis).

$$H_0 : \text{The defendant is innocent.}$$
$$H_A : \text{The defendant is guilty.}$$

Notice that if a jury finds a defendant *not guilty*, this does not necessarily mean the jury is confident in the person's innocence. They are simply not convinced of the alternative. This is also the case with hypothesis testing: **even if we fail to reject the null hypothesis, we do not accept the null hypothesis as truth.** Failing to find evidence in favor of the alternative is not the same as finding evidence that the null hypothesis is true.

7.3 Case study: sex discrimination

We consider a study investigating sex discrimination in the 1970s, set in the context of personnel decisions within a bank. The research question: "Are individuals who identify as female discriminated against in promotion decisions made by their managers who identify as male?" (Rosen and Jerdee 1974)

A note on the data. This 1972 study recorded each participant's sex as one of two categories, male or female, as was standard practice at the time. We keep that framing here to stay faithful to the original data. Sex and gender are now understood to be more varied than a single binary variable, and a study designed today might record them differently.

7.3.1 Observed data

The subjects were 48 male bank supervisors attending a management institute at the University of North Carolina in 1972. Each was asked to play the role of a bank's personnel director: review one personnel file and decide whether that candidate should be promoted to branch manager. The files were identical in every detail but one — half described a candidate identified as male, half a candidate identified as female. The researchers **randomly assigned** 24 supervisors to review a "male" file and the other 24 to review a "female" file.

Is this an observational study or an experiment? Who was randomly assigned, and what can the study tell us?

Show answer

It is an **experiment**: the researchers randomly assigned each supervisor to review a "male" or a "female" file (the files were otherwise identical). The **treatment** is the sex stated on the file. Because the assignment was random, the study can tell us whether the file's stated sex *caused* a difference in promotion decisions.

The results are summarized below. Each row is the sex stated on the *file* the supervisor reviewed (not the supervisor's own sex), and "Promoted" means the supervisor recommended promotion.

	Promoted	Not promoted	Total
Male	21	3	24
Female	14	10	24
Total	35	13	48

Why might these results look like discrimination — and why should we check chance before concluding that?

The promotion rate was much higher for "male" files (87.5%) than for "female" files (58.3%), though the files were otherwise identical — that gap looks like discrimination. But suppose the sex on the file made no difference to any supervisor. Even then, the random assignment could have happened to place more of the promotion-inclined supervisors in the group reviewing "male" files. Before concluding discrimination, we ask whether chance alone could reasonably produce a gap this large.

The observed difference in promotion rates is:

$$\hat{p}_M - \hat{p}_F = \frac{21}{24} - \frac{14}{24} = 0.875 - 0.583 = 0.292$$

We always subtract the "female" rate from the "male" rate, so a positive value means a higher promotion rate for "male" files — here, **29.2 percentage points higher**. The estimate is large but the sample is small, so it isn't yet clear whether it reflects real discrimination or ordinary chance. We label two competing claims:

- H_0: **Null hypothesis.** The variables `sex` and `decision` are independent — the sex on the file has no effect on the decision. The gap of 29.2 percentage points is just natural variability from the random assignment.
- H_A: **Alternative hypothesis.** `sex` and `decision` are *not* independent — otherwise-identical candidates are less likely to be promoted when the file states female than when it states male.

7.4 Building a randomization distribution

We want to know whether a 29.2-point gap would be surprising *if the null hypothesis were true*. So we simulate what could happen when the sex on the file has no effect on the decision, and see how the observed gap compares.

7.4.1 The card-shuffling metaphor

Picture the study as 48 cards — one for each supervisor's review of a single file. A card's **color** records the decision that supervisor made:

- 35 **red** cards — recommended promotion;
- 13 **white** cards — did not.

Two piles stand for the two file versions — a "male" pile and a "female" pile, each holding 24 cards, matching the original group sizes. Under the null hypothesis, the sex on the file has no effect, so a card's color has nothing to do with which pile it lands in.

To run one simulation:

1. **Shuffle** all 48 cards together.
2. **Deal** the cards into the two piles.
3. **Count** the red cards in each pile and divide by 24 to get each pile's promotion rate.
4. **Subtract** the "female" rate from the "male" rate.

Why does shuffling simulate "no effect"?

Under H_0, each supervisor would have made the *same* decision no matter which file they received — so we keep every card's color fixed. Dealing the cards out at random stands in for running the study's random assignment again. A pile can end up with more red cards purely by luck, so the two promotion rates can differ even though we built the simulation assuming the file's sex makes no difference.

Each shuffle keeps the totals fixed — 35 promoted, 13 not, and 24 cards per pile — and produces **one** difference in promotion rates *under the null hypothesis*.

7.4.2 One simulation

Suppose one shuffle and deal gives this result:

	Promoted	Not promoted	Total
Male	18	6	24
Female	17	7	24
Total	35	13	48

Find the simulated difference in promotion rates (the "male" rate minus the "female" rate). How does it compare with the observed 29.2 percentage points?

Show answer

The simulated difference is $18/24 - 17/24 \approx 0.042$, about 4.2 percentage points in favor of "male" files — far smaller than the observed 29.2 points. It came purely from dealing the cards at random while holding each supervisor's decision fixed.

This is only one possible result under H_0.

7.4.3 Many simulations

To judge how unusual the observed 29.2 points really is, we repeat the shuffle-and-deal many times, building up a picture of what *chance alone* looks like. Using a computer, we can do this hundreds or thousands of times.

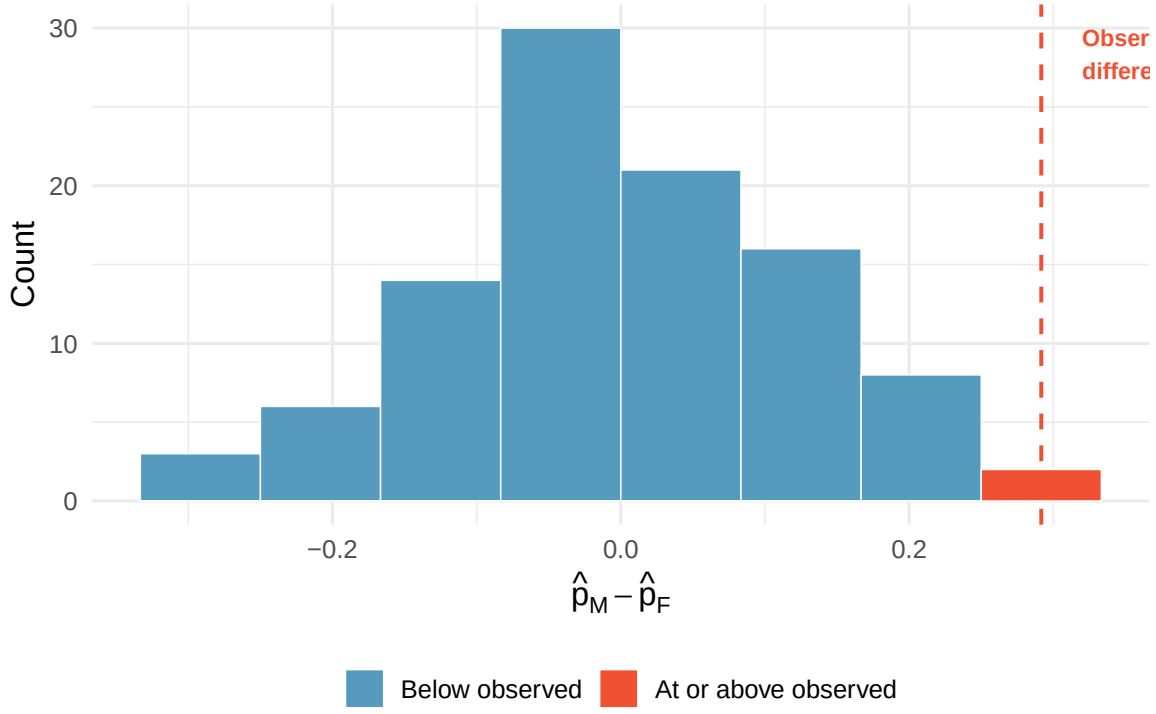


Figure 7.1: A histogram of differences from 100 simulations produced under the null hypothesis. The simulated sex and decision are independent. Differences at least as large as the observed 0.292 are highlighted in red. ↗ Try this live — shuffle the labels yourself and rebuild the null distribution.

The distribution of these simulated differences is centered around 0. This makes sense: under the null hypothesis, which makes no distinction between male and female files, we would expect the difference to be near zero with some random fluctuation.

How often would you observe a difference of at least 29.2 percentage points according to the randomization distribution? Often, sometimes, rarely, or never?

It appears that a difference of at least 29.2 percentage points under the null hypothesis would only happen about 2% of the time. Such a low probability indicates that observing such a large difference from chance alone is rare.

7.4.4 The conclusion

The difference of 29.2 percentage points is a rare event if there really is no effect of sex on promotion decisions. This provides us with two possible interpretations:

- If H_0 is true: Sex has no effect on promotion decisions, and we observed a difference that is so large it would only happen rarely by chance.
- If H_A is true: Sex has an effect on promotion decisions, and what we observed was actually due to equally qualified female candidates being discriminated against.

When we conduct formal studies, we reject a null position if the data strongly conflict with it. Since there was only about a 2% probability of obtaining a difference at least as large as 29.2 percentage points under the null hypothesis, we conclude that the data provide strong evidence of sex discrimination. We reject the null hypothesis in favor of the alternative.

Predict → Do → Explain. Open the Randomization Test Walkthrough — a gated activity on the sex-discrimination data. It starts from the observed data and the null hypothesis, then asks you to **predict** how big one shuffle's difference should be. You build the null distribution yourself with the **+1**, **+100**, and **+1000** buttons — each adds a simulated difference to the chart — with a reality-check gate at each stage. It closes with an **Explain** step where you put in your own words why the shuffle centers at zero, why the tail fraction *is* the p-value, and why a different random seed gives a slightly different estimate. Your predictions and explanation save locally in the tool, so you can return to your work on the same device.

> Tip
>
> **Open sandbox** (no scaffolding). Prefer to run the test yourself with your own data? Launch the Randomization Test: Difference in Proportions directly, enter the sex-discrimination table (or your own), generate 1,000 shuffled differences, and read off the tail fraction as the p-value.

7.5 Case study: opportunity cost

How rational and consistent is the behavior of the typical American college student? Here, we explore whether reminding students that money not spent now can be spent later causes them to be thriftier.

One hundred and fifty students at a university in the southwestern United States were recruited for the study (Frederick et al. 2009). All were given a scenario about purchasing a video on sale for $14.99, and asked to choose between two options. Option (A) — "Buy this entertaining video." — was the same for everyone. The two groups differed in *only* one place, option (B):

> **Control group** (B): Not buy this entertaining video.
>
> **Treatment group** (B): Not buy this entertaining video. **Keep the $14.99 for other purchases.**

The treatment group's added clause is a subtle nudge toward *opportunity cost* — the reminder that money not spent now can be spent later.

The hypotheses are:

- H_0: Reminding students that they can save money for later purchases will not have any impact on their spending decisions.
- H_A: Reminding students that they can save money for later purchases will reduce the chance they will continue with a purchase.

7.5.1 Observed data

The results are summarized below:

	Buy video	Not buy video	Total
Control	56	19	75
Treatment	41	34	75
Total	97	53	150

We define a "success" as a student who chooses *not* to buy the video. The difference in the proportion choosing not to buy is:

$$\hat{p}_T - \hat{p}_C = \frac{34}{75} - \frac{19}{75} = 0.453 - 0.253 = 0.200$$

The proportion of students who chose not to buy was 20 percentage points higher in the treatment group. Is this difference of 20 percentage points so large that it is unlikely to have occurred from chance alone?

7.5.2 Simulating under the null

Using the same technique, picture 150 cards — one for each student's decision. A card's **color** records that decision: 53 red cards mean "did not buy" (our "success") and 97 white cards mean "bought." Two stacks stand for the treatment and control groups, 75 cards each. Under the null hypothesis the reminder changes no one's decision, so each card's color stays fixed — we shuffle all 150 cards and deal them into the two stacks.

If we shuffle and deal the cards into simulated treatment and control stacks, how many red ("did not buy") cards would we expect in each stack? What would be the expected difference in proportions?

Since the stacks are equal in size, we would expect about $53/2 \approx 26$ or 27 red cards in each, yielding a difference of roughly 0 percentage points. Due to random chance, we might sometimes observe a little more or fewer than 26–27.

After running 1,000 simulations, the results show that the distribution of simulated differences is centered at 0, as expected under the null hypothesis.

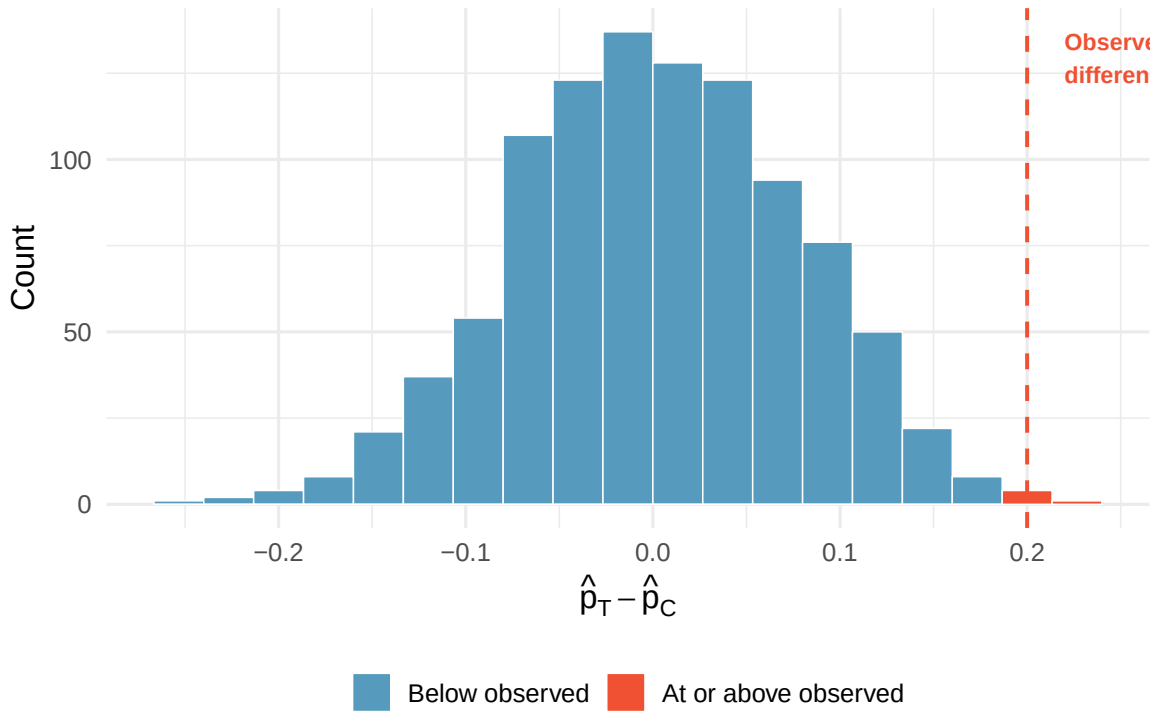


Figure 7.2: A histogram of 1,000 simulated differences produced under the null hypothesis. Differences at least as large as the observed 0.20 are highlighted in red. ↗ Try this live — run the randomization on your own data — vary reps and watch it settle.

Under the null hypothesis, we would observe a difference of at least +20 percentage points about 0.5% of the time (5 of the 1,000 shuffles). That is very rare! We conclude the data provide strong evidence there is a treatment effect: reminding students that they could spend money later lowers their chance of making a purchase. Because this is an experiment (with random assignment), we can make a causal claim.

7.6 The same idea, for means

Both case studies so far compared two groups on a **proportion**. But the shuffle never cared *what* we computed — it breaks the link between the group label and the outcome, whatever that outcome is. Here we run the exact same procedure on a **difference in means**.

A trainer wonders whether a short warm-up routine improves athletes' vertical jump. Twenty athletes are randomly assigned to two groups of 10 — one does the warm-up, the other does not — and each athlete's jump height (in inches) is measured.

The warm-up group averaged **24.05 inches** and the control group **22.71 inches**, an observed difference of

$$\bar{x}_{\text{warm-up}} - \bar{x}_{\text{control}} = 24.05 - 22.71 = 1.34 \text{ inches.}$$

Could a gap this large come from chance alone, if the warm-up made no real difference?

Under the null hypothesis the warm-up has no effect, so each athlete would have jumped the same height in either group — the label is just a tag. So we do exactly what we did before, changing only the last step:

1. **Shuffle** all 20 jump heights together.
2. **Deal** 10 into a "warm-up" pile and 10 into a "control" pile.
3. **Compute** the difference in *mean* jump height between the two piles.

The only change from the discrimination study is step 3 — we average the numbers on the cards instead of counting the red ones. Everything else, including *why* the shuffle is a fair stand-in for chance, is identical.

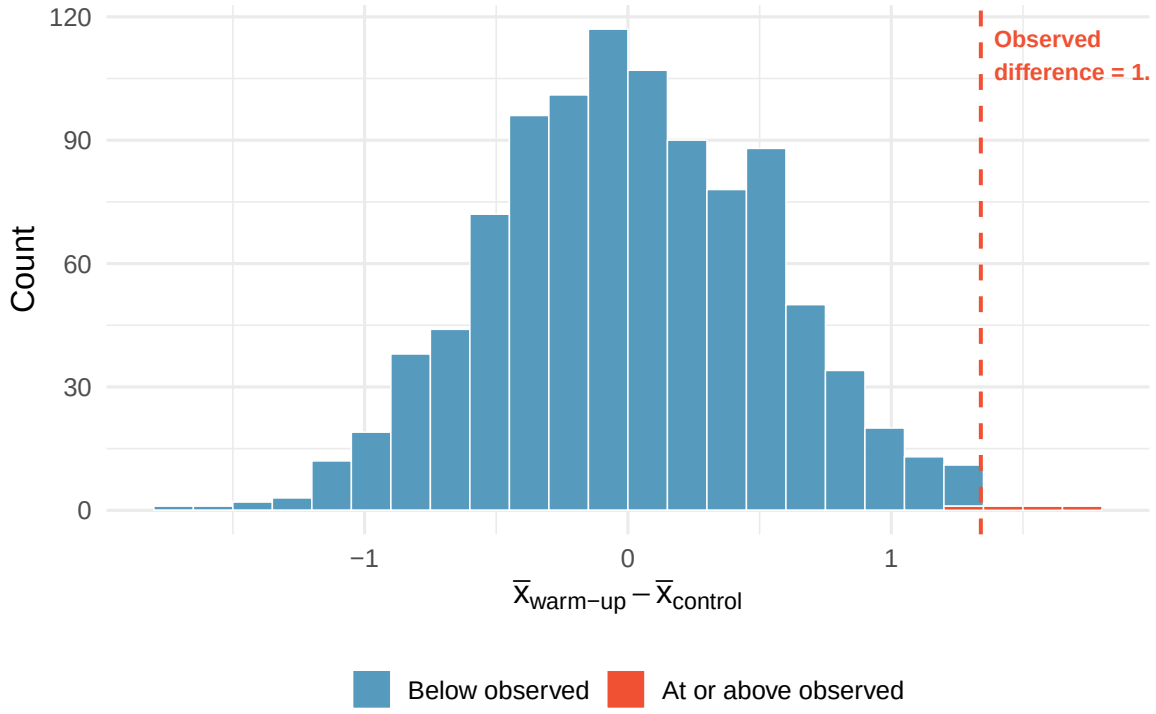


Figure 7.3: A histogram of 1,000 simulated differences in mean jump height under the null hypothesis that the warm-up has no effect. Shuffles at least as large as the observed 1.34 inches are shaded in red.

The simulated differences center on zero, exactly as before — under the null, the two groups should differ only by chance. The observed 1.34-inch gap sits far out in the right tail, reached by only a small fraction of shuffles, so the data provide evidence that the warm-up really does improve jump height.

Nothing about the shuffle-and-deal procedure changed between the discrimination study and this one except the statistic we computed at the end. What does that tell you about when a randomization test can be used?

Show answer

The same relabeling logic works for **any** statistic that compares two groups — a difference in proportions, a difference in means, or something else. As long as the null hypothesis says the group label is unrelated to the outcome, shuffling the labels and recomputing the statistic builds a valid null distribution.

7.7 p-values

The probability we computed in the previous sections — the chance of observing a result as extreme as or more extreme than the one actually observed, *if the null hypothesis were true* — has a name.

The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current dataset, if the null hypothesis were true. We typically use a **sample statistic** of the data, such as a difference in proportions, to compute the p-value. (Later, in the theory-based chapters, we standardize a sample statistic into a z- or t-score, and *that* standardized value is what we call the **test statistic**. In a randomization test we work directly with the sample statistic itself.)

In the sex discrimination study, the sample statistic was the difference in promotion rates. What was the sample statistic in the opportunity cost study?

The sample statistic in the opportunity cost study was the difference in the proportion of students who decided against the video purchase in the treatment and control groups: $\hat{p}_T - \hat{p}_C = 0.200$.

Interpreting the p-value.

- A **small p-value** (e.g., below 0.05) means that the observed result would be rare if the null hypothesis were true. This is evidence *against* the null hypothesis.
- A **large p-value** (e.g., above 0.10) means that the observed result is not particularly unusual under the null hypothesis. The data do not provide strong evidence against the null hypothesis.
- The p-value is **not** the probability that the null hypothesis is true. It is the probability of the observed data (or more extreme data) *given that* the null hypothesis is true.

A common mistake is to interpret the p-value as the probability that the null hypothesis is true (e.g., "there is a 2% chance that sex has no effect on promotions"). This is incorrect. The p-value tells us how *surprising* the data would be *if* the null hypothesis were true, not how likely the null hypothesis is to be true.

7.8 Statistical discernibility

We say that the data provide **statistically discernible** evidence against the null hypothesis if the p-value is less than a predetermined threshold called the **discernibility level**, denoted α (the Greek letter alpha).

> **i** A note on the vocabulary
>
> Traditionally this result has been called *statistically significant*, and the threshold α the *significance level* — terms you will still see in most other books. Because *significant* in everyday English suggests *large* or *important*, which a small but genuinely detectable effect may not be, we follow *Introduction to Modern Statistics* and instead call the result **statistically discernible** and the threshold the **discernibility level**.

The most commonly used discernibility level is $\alpha = 0.05$, meaning we reject the null hypothesis when the p-value is less than 5%. However, there is nothing magical about 0.05 — it is a convention, and in some fields or applications a stricter (e.g., 0.01) or more lenient (e.g., 0.10) threshold may be appropriate.

In the opportunity cost study, we found that a difference of at least 20 percentage points would occur only about 5-in-1,000 times if the reminder had no influence on student decisions. The p-value is approximately 0.005. Would you classify this as statistically discernible?

Yes. Since p-value $\approx 0.005 < 0.05 = \alpha$, the data provide statistically discernible evidence that US college students were actually influenced by the reminder.

What "statistically discernible" does and does not mean.

"Statistically discernible" means the p-value fell below the chosen discernibility level. It does **not** necessarily mean the effect is large or practically important. A very large sample can detect a tiny difference as "statistically discernible" even if the difference is too small to matter in practice.

Conversely, "not statistically discernible" does not mean there is no effect. It may simply mean the sample was too small to detect the effect.

7.8.1 Decision framework

The logic of hypothesis testing follows a consistent pattern:

If the p-value is...	Then we...	Because...
Less than α	**Reject** H_0	The data are unlikely under H_0; evidence supports H_A
Greater than or equal to α	**Fail to reject** H_0	The data are not sufficiently unusual under H_0

A researcher tests whether a new teaching method improves exam scores compared to the standard method. The p-value from the randomization test is 0.12. Using $\alpha = 0.05$, what conclusion should the researcher draw?

Show answer

Since p-value $= 0.12 > 0.05 = \alpha$, the researcher fails to reject the null hypothesis. The data do not provide sufficient evidence that the new teaching method improves exam scores. This does not mean the new method is ineffective — only that this study did not produce enough evidence to conclude it is effective.

7.9 Two-sided tests

In the sex discrimination study, we looked for differences in *one direction only* — whether female candidates were promoted at a lower rate. This is called a **one-sided** (or one-tailed) test. The alternative hypothesis specified a direction:

$$H_A : p_M - p_F > 0 \quad \text{(males promoted at a higher rate)}$$

Sometimes, however, we are interested in detecting a difference in *either direction*. For example, if we are studying whether a new drug changes the recovery rate compared to the standard treatment (it could raise it *or* lower it), we would use a **two-sided** (or two-tailed) test:

$$H_A : p_1 \neq p_2 \quad \text{(the proportions are different, in either direction)}$$

In a **one-sided test**, the alternative hypothesis specifies a direction (greater than or less than). The p-value is computed as the proportion of the randomization distribution in one tail.

In a **two-sided test**, the alternative hypothesis does not specify a direction (just "different"). The p-value is computed as the proportion of the randomization distribution in *both* tails — values at least as extreme as the observed statistic in either direction.

Suppose we are testing whether a coin is fair. We flip it 100 times and get 60 heads.

- **One-sided** ($H_A : p > 0.5$): The p-value counts only simulations with 60 or more heads.
- **Two-sided** ($H_A : p \neq 0.5$): The p-value counts simulations with 60 or more heads *and* simulations with 40 or fewer heads (equally extreme in the other direction).

For a symmetric randomization distribution, the two-sided p-value is approximately twice the one-sided p-value. If the one-sided p-value is 0.03, the two-sided p-value is approximately 0.06.

A researcher wants to know if a new fertilizer changes plant growth compared to no fertilizer. Should this be a one-sided or two-sided test? Write the hypotheses using appropriate notation, where μ_T is the mean growth with fertilizer and μ_C is the mean growth without.

Show answer

This should be a two-sided test because the fertilizer could either increase or decrease growth (perhaps it has a toxic effect in some soils). The hypotheses are: $H_0 : \mu_T - \mu_C = 0$ (no difference in mean growth) and $H_A : \mu_T - \mu_C \neq 0$ (the fertilizer changes mean growth, in either direction).

Choosing one-sided vs. two-sided.

- Use a **one-sided test** when you have a specific directional hypothesis *before looking at the data*. For example, "we believe this drug *lowers* blood pressure."
- Use a **two-sided test** when you want to detect a difference in either direction, or when you are not sure which direction the difference might go.
- When in doubt, use a two-sided test. It is the more conservative choice and is less susceptible to bias from choosing the direction after seeing the data.

7.10 The randomization test procedure: a summary

We can summarize the randomization test procedure as follows:

1. **Frame the research question in terms of hypotheses.** The null hypothesis (H_0) usually represents no difference or no effect. The alternative hypothesis (H_A) represents the claim of interest.
2. **Collect data.** If the research question focuses on associations but not causation, use an observational study. If you want to establish a causal connection, use an experiment with random assignment.
3. **Model what would happen under the null hypothesis.** Shuffle (randomize) the data to break any association between the explanatory and response variables. Compute the sample statistic for each shuffle. Repeat many times to build the **randomization distribution** (also called the null distribution).
4. **Compute the p-value.** Determine what proportion of the randomization distribution is at least as extreme as the observed sample statistic.
5. **Draw a conclusion.** If the p-value is small (less than α), reject H_0 and conclude there is evidence for H_A. If the p-value is not small, fail to reject H_0. Write the conclusion in context, in plain language.

See it in action. Open the Randomization Test: Difference in Proportions to run the full procedure on your own data — and watch how the p-value shifts when you change the sample size or the magnitude of the observed difference.

> **i** Testing a single proportion by simulation
>
> This chapter focused on comparing two groups because that is the most common use of a randomization test. The same simulate-under-H_0 logic also works for a *single* sample against a claimed proportion — instead of shuffling group labels, you draw new samples from a modeled null population. That one-sample simulation test is developed in the inference for proportions chapter (see the "Building the null distribution by simulation" section).

7.11 Chapter review

7.11.1 Summary

- A **hypothesis test** evaluates two competing claims: the null hypothesis (H_0), which represents no effect or no difference, and the alternative hypothesis (H_A), which represents the claim under investigation.
- A **randomization test** (or permutation test) assesses the null hypothesis by shuffling the data to simulate what would happen if there were truly no effect. The collection of simulated sample statistics forms the **randomization distribution**.
- The **sample statistic** is the summary value computed from the data (e.g., a difference in proportions) that we use to evaluate the hypotheses.
- The **p-value** is the probability of observing a sample statistic at least as extreme as the one observed, assuming the null hypothesis is true. It quantifies the strength of the evidence against H_0.
- We reject H_0 when the p-value is less than the **discernibility level** α (commonly 0.05). This is called a **statistically discernible** result.
- A **one-sided test** looks for effects in a specified direction; a **two-sided test** looks for effects in either direction.
- Failing to reject H_0 is not the same as proving H_0 is true.
- Statistical discernibility does not imply practical importance.

7.11.2 Key terms

Term	Definition
Hypothesis test	A formal procedure for evaluating competing claims using data
Null hypothesis (H_0)	The claim of no effect or no difference; the skeptical position
Alternative hypothesis (H_A)	The claim that there is an effect or difference
Randomization test	A hypothesis test that builds a null distribution by shuffling data
Randomization distribution	The distribution of the sample statistic under repeated randomizations
Sample statistic	The summary value computed from the data (e.g., a difference in proportions) used to evaluate the hypotheses
p-value	Probability of observing a result as extreme as or more extreme than the data, if H_0 is true
Discernibility level (α)	The threshold below which the p-value leads us to reject H_0
Statistically discernible	A result whose p-value falls below α

Term	Definition
One-sided test	Alternative hypothesis specifies a direction
Two-sided test	Alternative hypothesis does not specify a direction

7.12 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Hypotheses.** For each of the research statements below, note whether it represents a null hypothesis claim or an alternative hypothesis claim.

a. The number of hours that grade-school children spend doing homework predicts their future success on standardized tests.

b. King cheetahs on average run the same speed as standard spotted cheetahs.

c. For a particular student, the probability of correctly answering a 5-option multiple choice test is larger than 0.2 (i.e., better than guessing).

d. The mean length of African elephant tusks has changed over the last 100 years.

e. The risk of facial clefts is equal for babies born to mothers who take folic acid supplements compared with those from mothers who do not.

f. Caffeine intake during pregnancy affects mean birth weight.

g. The probability of getting in a car accident is the same if using a cell phone than if not using a cell phone.

2. **True null hypothesis.** Unbeknownst to you, let's say that the null hypothesis is actually true in the population. You plan to run a study anyway.

a. If the level of discernibility you choose (i.e., the cutoff for your p-value) is 0.05, how likely is it that you will mistakenly reject the null hypothesis?

b. If the level of discernibility you choose (i.e., the cutoff for your p-value) is 0.01, how likely is it that you will mistakenly reject the null hypothesis?

c. If the level of discernibility you choose (i.e., the cutoff for your p-value) is 0.10, how likely is it that you will mistakenly reject the null hypothesis?

3. **Identify hypotheses, I.** Write the null and alternative hypotheses in words and then symbols for each of the following situations.

a. New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Do these data provide convincing evidence that New Yorkers on average sleep less than 8 hours a night?

b. Employers at a firm are worried about the effect of March Madness, a basketball championship held each spring in the US, on employee productivity. They estimate that on a regular business day employees spend on average 15 minutes of company time checking personal email, making personal phone calls, etc. They also collect data on how much company time employees spend on such non- business activities during March Madness. They want to determine if these data provide convincing evidence that employee productivity decreases during March Madness.

4. **Identify hypotheses, II.** Write the null and alternative hypotheses in words and using symbols for each of the following situations.

a. Since 2008, chain restaurants in California have been required to display calorie counts of each menu item. Prior to menus displaying calorie counts, the average calorie intake of diners at a restaurant was 1100 calories. After calorie counts started to be displayed on menus, a nutritionist collected data on the number of calories consumed at this restaurant from a random sample of diners. Do these data provide convincing evidence of a difference in the average calorie intake of a diners at this restaurant?

b. Based on the performance of those who took the GRE exam between July 1, 2004 and June 30, 2007, the average Verbal Reasoning score was calculated to be 462. In 2021 the average verbal score was slightly higher. Do these data provide convincing evidence that the average GRE Verbal Reasoning score has changed since 2021?

5. **Side effects of Avandia.** Rosiglitazone is the active ingredient in the controversial type 2 diabetes medicine Avandia and has been linked to an increased risk of serious cardiovascular problems such as stroke, heart failure, and death. A common alternative treatment is Pioglitazone, the active ingredient in a diabetes medicine called Actos. In a nationwide retrospective observational study of 227,571 Medicare beneficiaries aged 65 years or older, it was found that 2,593 of the 67,593 patients using Rosiglitazone and 5,386 of the 159,978 using Pioglitazone had serious cardiovascular problems. These data are summarized in the contingency table below. (Graham et al. 2010)

Treatment	No	Yes	Total
Pioglitazone	154,592	5,386	159,978
Rosiglitazone	65,000	2,593	67,593
Total	219,592	7,979	227,571

See the next page for the questions.

a. Determine if each of the following statements is true or false. If false, explain why. *Be careful:* The reasoning may be wrong even if the statement's conclusion is correct. In such cases, the statement should be considered false.

   i. Since more patients on Pioglitazone had cardiovascular problems (5,386 vs. 2,593), we can conclude that the rate of cardiovascular problems for those on a Pioglitazone treatment is higher.

   ii. The data suggest that diabetic patients who are taking Rosiglitazone are more likely to have cardiovascular problems since the rate of incidence was (2,593 / 67,593 = 0.038) 3.8% for patients on this treatment, while it was only (5,386 / 159,978 = 0.034) 3.4% for patients on Pioglitazone.

   iii. The fact that the rate of incidence is higher for the Rosiglitazone group proves that Rosiglitazone causes serious cardiovascular problems.

   iv. Based on the information provided so far, we cannot tell if the difference between the rates of incidences is due to a relationship between the two variables or due to chance.

b. What proportion of all patients had cardiovascular problems?

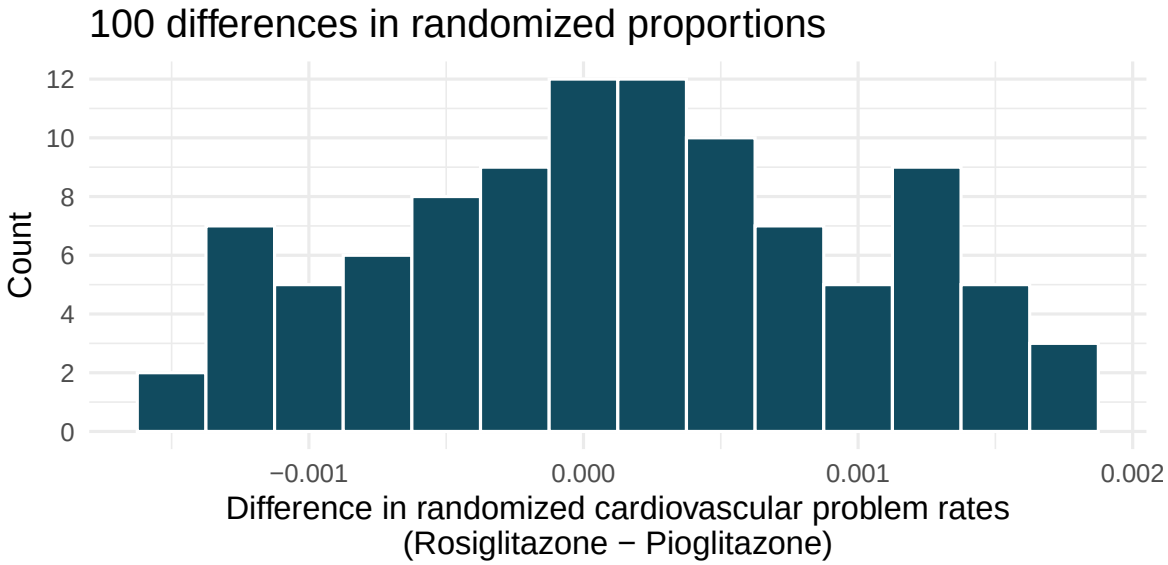
c. If the type of treatment and having cardiovascular problems were independent, how many patients in the Rosiglitazone group would we expect to have had cardiovascular problems?

d. We can investigate the relationship between outcome and treatment in this study using a randomization technique. While in reality we would carry out the simulations required for randomization using statistical software, suppose we actually simulate using index cards. In order to simulate from the independence model, which states that the outcomes were independent of the treatment, we write whether each patient had a cardiovascular problem on cards, shuffled all the cards together, then deal them into two groups of size 67,593 and 159,978. We repeat this simulation 100 times and each time record the difference between the proportions of cards that say "Yes" in the Rosiglitazone and Pioglitazone groups. Use the histogram of these differences in proportions to answer the following questions.

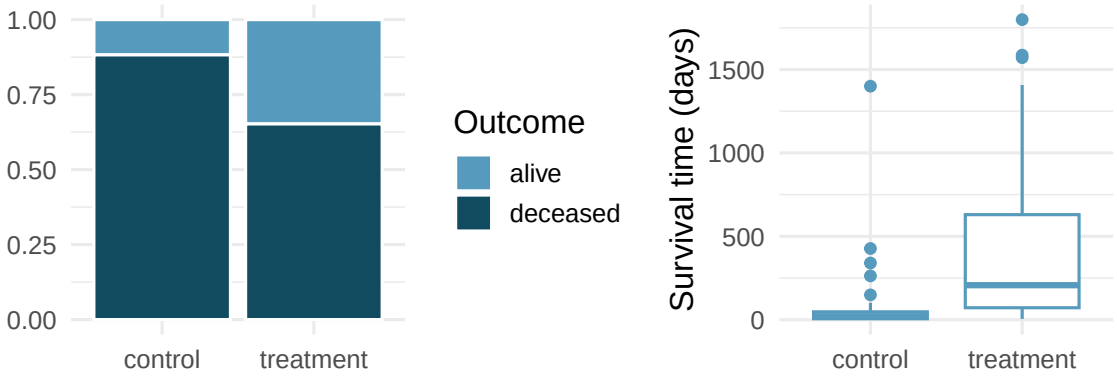
   i. What are the claims being tested?

   ii. Compared to the number calculated in part (b), which would provide more support for the alternative hypothesis, *higher* or *lower* proportion of patients with cardiovascular problems in the Rosiglitazone group?

   iii. What do the simulation results suggest about the relationship between taking Rosiglitazone and having cardiovascular problems in diabetic patients?


100 differences in randomized proportions
Count
12
10
8
6
4
2
0
−0.001
0.000
0.001
0.002
Difference in randomized cardiovascular problem rates
(Rosiglitazone − Pioglitazone)

6. **Heart transplants.** The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that they were gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable `transplant` indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Of the 34 patients in the control group, 30 died. Of the 69 people in the treatment group, 45 died. Another variable called `survived` was used to indicate whether the patient was alive at the end of the study. (Turnbull, Brown, and Hu 1974)



a. Does the stacked bar plot indicate that survival is independent of whether the patient got a transplant? Explain your reasoning.

b. What do the box plots suggest about the efficacy of heart transplants.

c. What proportions of patients in the treatment and control groups died?

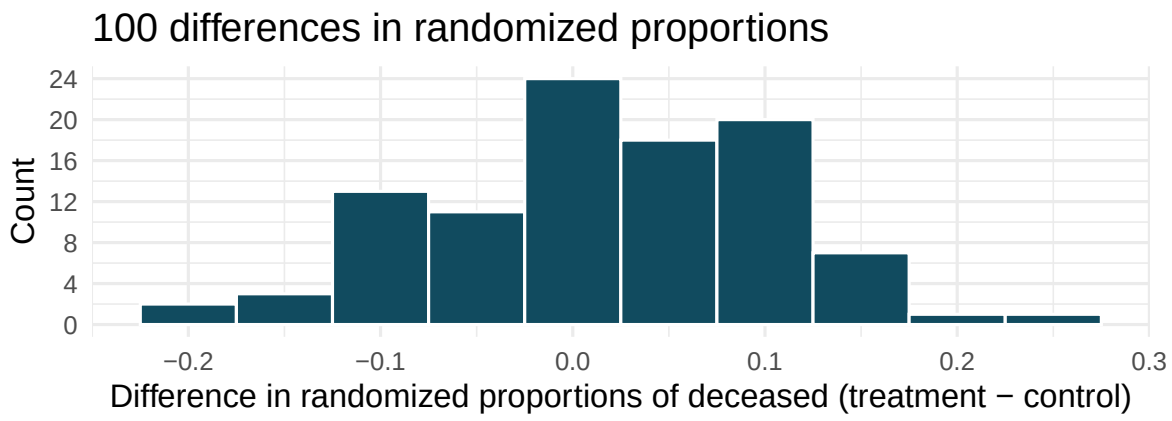
d. One approach for investigating whether the treatment is discernably effective is randomization testing.

   i. What are the claims being tested?

   ii. The paragraph below describes the set up for a randomization test, if we were to do it without using statistical software. Fill in the blanks with a number or phrase.

   We write *alive* on _____ cards representing patients who were alive at the end of the study, and *deceased* on _____ cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size _____ representing treatment, and another group of size _____ representing control. We calculate the difference between the proportion of *deceased* cards in the treatment and control groups (treatment - control) and record this value. We repeat this 100 times to build a distribution centered at _____. Lastly, we calculate the proportion of simulations where the simulated differences in proportions are _____. If this proportion is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.

   iii. What do the simulation results shown below suggest about the effectiveness of heart transplants?



Datasets: `avandia` (openintro) | `heart_transplant` (openintro)

StatLens Exercises

These exercises focus on the parts of a randomization test a calculator *can't* do for you: **framing the question**, **picking the right shuffling scheme**, **reading the null distribution**, and **interpreting a p-value** without falling into the standard traps. Let StatLens do the arithmetic.

1. **Frame the question.** A trial randomly assigns 90 students to either an online tutoring program or a control group, then measures their final exam scores. The mean score in the tutoring group is 4.2 points higher than in the control group.

   a. What is the parameter of interest? Describe it in words.
   b. Write H_0 and H_A for testing whether tutoring *improves* exam scores. Is the test one-sided or two-sided, and why?
   c. The randomization test will simulate the "no effect" world by **shuffling group labels** across students. *Why is shuffling labels the right way to simulate H_0 here, and what feature of the study design makes shuffling labels valid?*

2. **Choose the procedure.** For each scenario, name the randomization shuffle you would use — **shuffle group labels** (independent groups), **shuffle paired-difference signs** (paired data), **shuffle the values of one variable** (correlation/independence), or **a Bernoulli null simulation** (one proportion vs. a fixed value).

   a. Did a coin biased toward heads come up "heads" more than 50% of the time in 100 tosses?
   b. Are male and female job applicants offered different mean starting salaries?
   c. Do students score higher on a test taken **after** a study skills workshop compared to **before**?
   d. Is hours of TV per week associated with hours of sleep per night?
   e. Do four different fertilizer brands produce different mean yields?

3. **Read the null distribution.** Open the Randomization Test — Difference in Proportions (the gated sex-discrimination walkthrough). Step through it until you have the full null distribution (the walkthrough's **+100** and **+1000** steps, about 1,100 shuffled differences) in front of you.

   a. Where is the null distribution centered? Why does it have to be centered there, by construction?
   b. The observed sample difference is marked on the dotplot. Roughly what proportion of the *shuffled* differences are at least as extreme as the observed one (one-sided)? That proportion *is* the p-value.
   c. Suppose the observed difference had landed *near* the middle of the null distribution instead. What would the p-value be approximately, and what would you conclude?

4. **Run it and interpret it (means).** Open the Randomization Test for Difference in Means with any two-group numeric dataset.
   a. Run 1000 shuffles and read off the p-value. State your decision at $\alpha = 0.05$.
   b. Write a one-sentence conclusion **in context**.
   c. Re-run the test with only 100 shuffles, then with 10000 shuffles. How does the p-value change? Why do we usually report a number that comes from "more shuffles, not fewer"?
   d. Which statement is the correct interpretation of a p-value of 0.03?
      i. There is a 3% probability that the null hypothesis is true.
      ii. If the two groups really had the same mean, we would see a difference at least as extreme as ours in about 3% of randomly assigned studies.
      iii. The probability that the alternative hypothesis is true is 0.97.
5. **Simulation as a sanity check on the normal approximation.** A textbook says: "When the sample sizes are large and the data are not too skewed, the randomization distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal." You will check this by eye.
   a. Open any moderate-n two-group numeric dataset in the randomization test tool and run 5000 shuffles. Describe the **shape**, **center**, and **spread** of the resulting null distribution.
   b. Why does the *spread* of the null distribution tell you the *standard error* of $\bar{x}_1 - \bar{x}_2$ — and why does that mean a randomization test essentially "discovers" the same SE the t-formula computes from s_1, s_2, n_1, n_2?
   c. Now try the same on a *small*, skewed dataset (say $n_1 = n_2 = 6$, with one outlier). Is the null distribution still approximately normal? If you ran a two-sample t-test on this same data, would the t distribution be a good match for the randomization distribution? Where would they disagree?
   d. A student claims that the randomization test "only works for large samples." Why is this exactly backwards — which test works *better* in small samples, the t-test or the randomization test, and why?

Chapter 8

Decision Errors

In the previous chapter we built the logic of a hypothesis test: state a null and an alternative, measure the evidence with a p-value, and reject H_0 when the p-value falls below the discernibility level α. But a decision based on data can be **wrong** in two different ways. This chapter names those two mistakes — the **Type I error** and the **Type II error** — organizes them in a decision table, gives each one an error rate (α and β), and examines the tradeoff between them. Naming these errors is what lets us choose a discernibility level deliberately rather than out of habit.

StatLens: Decision Errors simulator. Open the Decision Errors simulator (or the open sandbox). It uses a plain one-proportion framing — H_0: $p = 0.50$ vs H_A: $p > 0.50$ — and lets you decide whether the alternative is *actually* true (and at what success rate), pick n and α, then run many studies. Each study is one mark: a filled blue dot if the test detected a real effect, a hollow dot if it missed (Type II), an orange □ if it false-alarmed under a true H_0 (Type I), and a light gray dot if it correctly kept a true H_0. The hit/miss/false-alarm strip is the picture this chapter is about. (*Power gets a fuller treatment in Chapters 29–31; the analytical Power & Error Visualizer is there if you want a quick preview.*)

8.1 Type I error

Recall the structure of a hypothesis test. We weigh two competing claims:

- H_0 (the **null hypothesis**): there is no effect, no difference, or no relationship.
- H_A (the **alternative hypothesis**): there *is* an effect, difference, or relationship.

We collect data, compute a test statistic, and find a p-value — the probability of observing results at least as extreme as ours if H_0 were true. If the p-value is below our chosen discernibility level α, we reject H_0.

But rejecting H_0 when it is actually true is a mistake, and that mistake has a name.

A **Type I error** occurs when we reject the null hypothesis H_0 even though H_0 is actually true. In other words, we conclude there is an effect when in reality there is none.

The probability of a Type I error is controlled by the discernibility level α. When we set $\alpha = 0.05$, we are accepting a 5% chance of making this kind of mistake when H_0 is true — declaring a "discovery" that is really just noise. In the sex discrimination study of the previous chapter, a Type I error would mean concluding that the managers' decisions depended on the candidate's sex when, in truth, they did not.

8.2 Type II error: missing a real effect

Suppose a new medication truly does lower blood pressure compared to a placebo. A clinical trial is conducted, data are collected, and the researchers perform a hypothesis test. But the p-value comes back at 0.14 — not small enough to reject H_0 at the $\alpha = 0.05$ level. The researchers conclude: "We did not find sufficient evidence that the medication works."

But the medication *does* work. The study simply failed to detect the real effect. This is a **Type II error**.

A **Type II error** occurs when we fail to reject the null hypothesis H_0 even though the alternative hypothesis H_A is actually true. In other words, there is a real effect, but our test misses it.

Type II errors are sometimes called "false negatives" — the test comes back negative (fail to reject H_0) when the truth is positive (the effect exists). In medical screening, this is analogous to a test that says "no disease" when the patient actually has the disease. In a court system, it is analogous to acquitting a guilty person.

8.3 The decision table

Every hypothesis test ends with one of two decisions: reject H_0 or fail to reject H_0. And the truth is one of two realities: H_0 is true or H_A is true. Combining these gives four scenarios:

Table 8.1: Four possible outcomes of a hypothesis test.

	H_0 is true	H_A is true
Reject H_0	Type I error	Correct decision
Fail to reject H_0	Correct decision	Type II error

Two of the four outcomes are correct decisions:

- Rejecting H_0 when H_A is true — we correctly detected the real effect.
- Failing to reject H_0 when H_0 is true — we correctly avoided a false alarm.

The other two outcomes are errors:

- **Type I error** (upper-left): We rejected H_0 but it was actually true. False alarm.
- **Type II error** (lower-right): We failed to reject H_0 but H_A was actually true. Missed detection.

In a US court, the defendant is either innocent (H_0) or guilty (H_A). What does a Type I error represent in this context? What does a Type II error represent? Table 8.1 may be useful.

If the court makes a Type I error, this means the defendant is innocent (H_0 true) but wrongly convicted. A Type II error means the court failed to reject H_0 (i.e., failed to convict the person) when they were in fact guilty (H_A true).

A pharmaceutical company tests whether a new drug reduces cholesterol more than a placebo. They set up H_0: the drug has no effect, and H_A: the drug reduces cholesterol. Describe what a Type I error and a Type II error would mean in this context.

Show answer

A Type I error would mean concluding the drug reduces cholesterol when it actually does not — the company might invest millions in producing an ineffective drug. A Type II error would mean failing to detect a real cholesterol-lowering effect — the company might shelve a drug that actually works.

A quality control inspector tests whether a batch of light bulbs has a defect rate above 2%. She sets up H_0: defect rate ≤ 0.02 and H_A: defect rate > 0.02. Which error is worse in this context: Type I or Type II? Why?

Show answer

A Type II error is arguably worse here: the inspector would fail to detect that the defect rate is too high, and defective bulbs would be shipped to customers. A Type I error means rejecting a good batch, which wastes product but doesn't harm customers. The relative costs depend on the specific situation, but in quality control, missing a real problem (Type II) often has more serious consequences.

8.4 Error rates: α and β

We give the probabilities of the two error types their own Greek-letter names:

The **Type I error rate**, denoted α, is the probability of rejecting H_0 when H_0 is true:

$$\alpha = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

The **Type II error rate**, denoted β, is the probability of failing to reject H_0 when H_A is true:

$$\beta = P(\text{fail to reject } H_0 \text{ when } H_A \text{ is true})$$

The discernibility level α is something we choose before the test — it is under our direct control. Common choices are $\alpha = 0.05$, $\alpha = 0.01$, and $\alpha = 0.10$.

The Type II error rate β, on the other hand, is **not** something we directly choose. It depends on several factors that we will explore shortly: the true effect size, the sample size, the variability in the data, and the chosen discernibility level α. Unlike α, β is usually unknown because we don't know the true state of the world — if we knew H_A were true and knew the exact effect size, we wouldn't need to run the test.

This asymmetry between α and β is a fundamental feature of the hypothesis testing framework. We control one error rate directly (α) and try to manage the other (β) through careful study design.

8.5 The tradeoff between α and β

How could we reduce the Type I error rate in US courts? What influence would this have on the Type II error rate?

To lower the Type I error rate, we might raise our standard for conviction from "beyond a reasonable doubt" to "beyond a conceivable doubt" so fewer people would be wrongly convicted. However, this would also make it more difficult to convict the people who are actually guilty, so we would make more Type II errors.

How could we reduce the Type II error rate in US courts? What influence would this have on the Type I error rate?

Show answer

To lower the Type II error rate, we want to convict more guilty people. We could lower the standards for conviction from "beyond a reasonable doubt" to "beyond a little doubt". Lowering the bar for guilt will also result in more wrongful convictions, raising the Type I error rate.

The example and guided practice above provide an important lesson: if we reduce how often we make one type of error, we generally make more of the other type.

This tradeoff is not just a courtroom analogy — it is baked into the mathematics of hypothesis testing. Here is the core intuition:

- **Lowering** α means we require stronger evidence (a smaller p-value) before rejecting H_0. This makes us less likely to reject H_0 when it is true (good), but also less likely to reject H_0 when it is false (bad). So β increases.
- **Raising** α means we accept weaker evidence for rejecting H_0. This makes us more likely to detect real effects (β decreases), but also more likely to make false discoveries (α increases).

The α-β tradeoff. For a fixed sample size and effect size, decreasing α increases β, and vice versa. You cannot reduce both error rates simultaneously without changing something else — typically the sample size.

To see why this makes sense, imagine a number line representing all possible values of a test statistic. Under H_0, the test statistic is centered at some null value. The rejection region is the set of values far enough from the null to trigger rejection. Making α smaller means shrinking the rejection region — fewer values lead to rejection. But if H_A is true, the test statistic is centered somewhere else. A smaller rejection region means less of the alternative distribution falls inside it, so we are less likely to reject — β goes up.

The only way to reduce *both* α and β at the same time is to collect more data. More data makes the test statistic more precise (smaller standard error), which makes it easier to distinguish between H_0 and H_A — allowing us to keep α small while also keeping β small. This is why sample size planning (the sample size planning chapter) is so important.

Predict → Do → Explain. Open the Decision Errors simulator — the gated activity opens with a **Predict** step (commit to how many of 500 studies will reject H_0 when H_0 is actually true, and what each such rejection is), walks you through both scenarios one step at a time — varying n, α, and the true success rate — and closes with an **Explain** step where you articulate the α–β tradeoff in your own words. Each mark in the tool is one study: **filled blue dot** = detected the effect (correct rejection), **hollow dot** = missed the effect (Type II), **orange** □ = false alarm under a true H_0 (Type I). Your predictions and explanation save locally in the tool, so you can leave the activity and return to your work on the same device.

8.6 Choosing a discernibility level

The **discernibility level** (traditionally called the **significance level**) provides the cutoff for the p-value which will lead to a decision of "reject the null hypothesis." The traditional level is 0.05. However, it is sometimes helpful to adjust the discernibility level based on the application. We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.

If making a Type I error is dangerous or especially costly, we should choose a small discernibility level (e.g., 0.01 or 0.001). If we want to be very cautious about rejecting the null hypothesis, we demand very strong evidence favoring the alternative H_A before we would reject H_0.

If a Type II error is relatively more dangerous or much more costly than a Type I error, then we should choose a higher discernibility level (e.g., 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

Discernibility levels should reflect consequences of errors. The discernibility level selected for a test should reflect the real-world consequences associated with making a Type I or Type II error.

A city water department tests whether the lead content in drinking water exceeds the safe limit. They set H_0: lead level $\leq$ safe limit, and H_A: lead level $>$ safe limit. Should they use $\alpha = 0.01$ or $\alpha = 0.10$?

A Type II error here — failing to detect dangerous lead levels — could have serious public health consequences. A Type I error — declaring the water unsafe when it is actually fine — would cause inconvenience and expense but is far less harmful than allowing people to drink contaminated water. Therefore, the water department should use a larger α (like 0.10) to make it easier to detect real contamination, accepting a slightly higher chance of false alarms.

A social media company wants to test whether a new recommendation algorithm increases average time spent on the platform. They set H_0: no difference, H_A: the new algorithm increases engagement. Should they use $\alpha = 0.01$ or $\alpha = 0.10$?

The stakes here are relatively low. A Type I error means rolling out an algorithm that doesn't actually help — some wasted development effort, and easy to undo. A Type II error means keeping the old algorithm when the new one would have been slightly better — a missed opportunity. Neither mistake is costly, and a false alarm can be caught cheaply by rerunning the test, so there is no reason to demand very strong evidence. The company should use $\alpha = 0.10$: it makes a real improvement easier to detect, and the extra false alarms are inexpensive. (A strict $\alpha = 0.01$ fits the opposite situation, where a false alarm is the expensive mistake.)

8.7 Looking ahead: the probability of a Type II error

We can choose α directly, but β — the Type II error rate — depends on things we do not control: the size of the real effect, the sample size, and the variability in the data. Quantifying β, and its complement the **power** of a test, is the subject of an optional later chapter on Type II error and power. There we use simulation to *see* how often a real effect slips past a test, and how planning the sample size brings the Type II error rate under control.

8.8 Chapter review

8.8.1 Summary

- Every hypothesis test ends in one of four outcomes, organized by the **decision table**: two correct decisions and two errors.
- A **Type I error** rejects a true H_0 (a false alarm); its probability is the discernibility level α, which we choose.
- A **Type II error** fails to reject H_0 when H_A is true (a missed detection); its probability is β, which we do not directly control.
- For a fixed sample size and effect size, lowering α raises β and vice versa — the α–β tradeoff. Collecting more data is the way to reduce both.
- The discernibility level should be chosen to reflect the real-world consequences of each kind of error.

8.8.2 Key terms

Term	Definition
Type I error	Rejecting H_0 when it is true (false positive)

Term	Definition
Type II error	Failing to reject H_0 when H_A is true (false negative)
Discernibility level (α)	The probability of a Type I error; chosen before the test
Type II error rate (β)	The probability of a Type II error
Decision table	The two-by-two table of test decision versus the truth
α–β tradeoff	At fixed n, lowering one error rate raises the other

8.9 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Testing for Fibromyalgia.** A patient named Diana was diagnosed with Fibromyalgia, a long-term syndrome of body pain, and was prescribed anti-depressants. Being the skeptic that she is, Diana didn't initially believe that anti-depressants would help her symptoms. However after a couple months of being on the medication she decides that the anti-depressants are working, because she feels like her symptoms are in fact getting better.

a. Write the hypotheses in words for Diana's skeptical position when she started taking the anti-depressants.

b. What is a Type I error in this context?

c. What is a Type II error in this context?

2. **Testing for food safety.** A food safety inspector is called upon to investigate a restaurant with a few customer reports of poor sanitation practices. The food safety inspector uses a hypothesis testing framework to evaluate whether regulations are not being met. If he decides the restaurant is in gross violation, its license to serve food will be revoked.

a. Write the hypotheses in words.

b. What is a Type I error in this context?

c. What is a Type II error in this context?

d. Which error is more problematic for the restaurant owner? Why?

e. Which error is more problematic for the diners? Why?

f. As a diner, would you prefer that the food safety inspector requires strong evidence or very strong evidence of health concerns before revoking a restaurant's license? Explain your reasoning.

3. **True / False.** Determine if the following statements are true or false, and explain your reasoning. If false, state how it could be corrected.

a. If a given value (for example, the null hypothesized value of a parameter) is within a 95% confidence interval, it will also be within a 99% confidence interval.

b. Decreasing the discernibility level (α) will increase the probability of making a Type I error.

c. Suppose the null hypothesis is $p = 0.5$ and we fail to reject H_0. Under this scenario, the true population proportion is 0.5.

d. With large sample sizes, even small differences between the null value and the observed point estimate, a difference often called the effect size, will be identified as statistically discernible.

4. **Online communication.** A study suggests that 60% of college student spend 10 or more hours per week communicating with others online. You believe that this is incorrect and decide to collect your own sample for a hypothesis test. You randomly sample 160 students from your dorm and find that 70% spent 10 or more hours a week communicating with others online. A friend of yours, who offers to help you with the hypothesis test, comes up with the following set of hypotheses. Indicate any errors you see.

$$H_0 : \hat{p} < 0.6 \qquad H_A : \hat{p} > 0.7$$

5. **Estimating π.** In a class activity, each of 100 students experimentally estimates the value of π, 10 separate times. Using the 10 measurements for π (10 values of $\hat{\pi}$), each student calculates a confidence interval for π. In grading the 100 student assignments, the professor marks 7 of the assignments wrong, indicating that the 7 students must have done their experiments or analysis incorrectly because each of the 7 students reported confidence intervals that did not capture the known true value of π, roughly 3.14159. Was the professor correct to mark the assignments wrong for having CIs that did not capture the value of 3.14159? Explain.

6. **Fermenting yeast.** Twenty students work individually in a biology lab to test whether using raw sucrose versus refined sugar will lead to the same yeast fermentation rate. Each student runs a full experiment independently of the other students in the lab. Of the twenty students, twelve are able to reject the null hypothesis and to claim that the fermentation rates are different.

a. Explain what type of error was likely to have occurred in this situation.
b. What change would you suggest that would lower the error rate?

7. **Hypothesis statements.** For each of the research claims below, fill in the value and the direction of the null and alternative hypotheses. That is, complete all aspects of the following hypothesis statements. Additionally, for each item, describe p in words.

$$H_0 : p\_\_\_\_\_ \qquad H_A : p\_\_\_\_\_$$

a. On a pre-test to assess knowledge of the upcoming material, a professor wants to determine if their students know, on average, more than if they were just randomly guessing. The pre-test is 30 multiple choice questions, where each question has 5 possible responses.

b. A standard treatment is known to reduce blood pressure in 32% of patients. A clinical trial is conducted to assess whether a new medical intervention will produce results which are different than the standard treatment, in terms of the percent of patients who will have reduced blood pressure.

c. In the last presidential election 67% of registered voters turned out to vote. Will the next presidential election have a higher turn-out of voters?

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Frame the question (and the two ways to be wrong).** A pharmaceutical company tests whether a new drug raises the success rate above the existing 50% standard.
   a. Write H_0 and H_A in symbols.
   b. In context, describe a **Type I error** and a **Type II error** in plain English (no Greek letters).
   c. In context, who pays the cost of each error — patients, the company, regulators? When the costs of the two errors are unequal, which one do you typically want to control more tightly, and how?

2. **Pick the right α for the situation.** For each scenario, would you set α at the **conventional** 0.05, **lower** (e.g., 0.01), or **higher** (e.g., 0.10)? Justify.
   a. A smoke alarm: H_0 = "no fire" vs. H_A = "fire." Type I = false alarm; Type II = silent during a real fire.
   b. Approving a new vaccine: H_0 = "no effect" vs. H_A = "effective."
   c. An A/B test of two webpage headlines: H_0 = "same click rate" vs. H_A = "different."
   d. A criminal trial: H_0 = "innocent" vs. H_A = "guilty."

3. **Simulate the error rates.** Open the Decision Errors simulator and step through the gated activity. Run the *truth-is-H_0* scenario first.
   a. Set truth = H_0 ($p = 0.50$), $n = 100$, $\alpha = 0.05$. Run 500 studies. Of those, roughly what fraction reject H_0? What error does each rejection represent?
   b. Drop α to 0.01 and re-run 500 studies. Did the fraction of rejections go up or down? Is this consistent with what α is *supposed* to be?
   c. Now flip truth to H_A with $p_{\text{true}} = 0.65$, $n = 100$, $\alpha = 0.05$. Run 500 studies. Roughly what fraction reject? Of the **non-rejections**, what kind of error are those?
   d. With truth = H_A, $p_{\text{true}} = 0.65$, $\alpha = 0.05$, run at $n = 50$, $n = 100$, and $n = 200$. How does the rejection rate (the **power**) change with n?

4. **The conditional nature of α and β.**
   a. Why is α a probability **conditional on H_0 being true**, and β a probability **conditional on a specific H_A**?
   b. A student says: "If $\alpha = 0.05$, then 5% of rejections are wrong." Why is this wrong, even though it sounds plausible?
   c. Suppose 1000 hypothesis tests are run, all at $\alpha = 0.05$. In 800 of them H_0 is actually true. How many *false positives* (Type I errors) do you expect? How does this differ from the "5% are wrong" framing?
   d. Connect to **multiple comparisons**: why does running 20 tests at $\alpha = 0.05$ each give you a substantial chance of at least one false rejection, even if all 20 null hypotheses are true?

5. **A glimpse of power (formal treatment in Ch 29–31).** Suppose a clinical trial will test whether a drug raises the success rate from 50% to 65%. With $n = 100$ patients and $\alpha = 0.05$ (one-sided).
   a. Predict, without running anything: do you think this trial has enough patients to *reliably detect* a real 15-percentage-point effect? "Reliably" usually means at least 80% power.
   b. Open the Decision Errors simulator with truth = H_A, $p_{\text{true}} = 0.65$, $n = 100$, $\alpha = 0.05$ and run 1000 studies. What is the simulated power?
   c. Given the power you found, what's the practical implication — if the drug really works, what's the chance the trial *fails*?
   d. Name **two** things you could change *before* the trial begins to **raise** its power (say, if the effect had turned out smaller and power fell short). Which one is usually under the researcher's

control?

Chapter 9

Bootstrap Confidence Intervals

In the sampling variability chapter, we saw that sample statistics vary from sample to sample. This variability is not a nuisance — it's the key to quantifying uncertainty. In this chapter, we learn **bootstrapping**: a simulation-based method for constructing confidence intervals that requires no formulas and no assumptions about the shape of the population.

9.1 The big idea: resampling

In 2004, researchers collected a sample of 648 pennies in circulation and recorded how old each penny was (2004 minus the year minted). The question: what was the average age of *all* pennies in circulation in the United States at that time?

The sample mean is $\bar{x} = 10.4$ years. This is our **point estimate** of the population mean μ. But how precise is this estimate? If we collected a different sample of 648 pennies, we'd get a different $\bar{x}$. We can't keep collecting new samples — we only have this one.

The bootstrap idea: Treat your sample as a stand-in for the population and resample *from it*. Each **bootstrap resample** is drawn with replacement from your original sample, at the same sample size n.

A **bootstrap resample** is a sample of size n drawn **with replacement** from the original sample. The statistic computed from a bootstrap resample is called a **bootstrap statistic**. The collection of many bootstrap statistics forms the **bootstrap distribution**.

Why does resampling from our own sample work? If the original sample is representative of the population, then resampling from it mimics what would happen if we could repeatedly sample from the population itself. The variability we see in the bootstrap distribution approximates the true sampling variability of our statistic.

9.2 Bootstrapping a sample mean

Let's build a bootstrap confidence interval step by step.

9.2.1 Step 1: Collect data

Our sample of 648 pennies has a right-skewed distribution — many young pennies and a long tail of older ones:

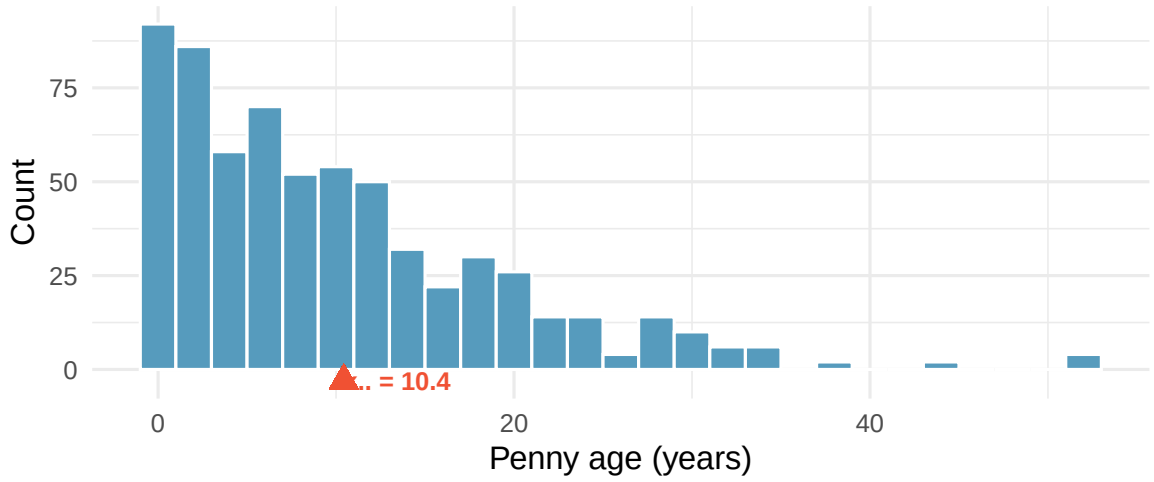


Figure 9.1: Distribution of ages for 648 pennies sampled from circulation. The distribution is right-skewed with a mean of about 10.4 years. The sample mean is marked with a red triangle. Open in StatLens

The sample mean is $\bar{x} = 10.4$ years. Notice the distribution is clearly not bell-shaped — it's skewed right. Bootstrapping doesn't care. It works regardless of the population shape.

9.2.2 Step 2: Resample with replacement

We draw a new sample of 648 values from the original 648, *with replacement*. Some pennies will appear more than once; others will be left out. Each such draw is one **bootstrap resample**, and the mean of that resample is a **bootstrap statistic** $\bar{x}^*$.

9.2.3 Step 3: Repeat many times

We repeat Step 2 a large number of times — typically $B = 1{,}000$ or more — and record the bootstrap mean $\bar{x}^*$ each time. This gives us a **bootstrap distribution**.

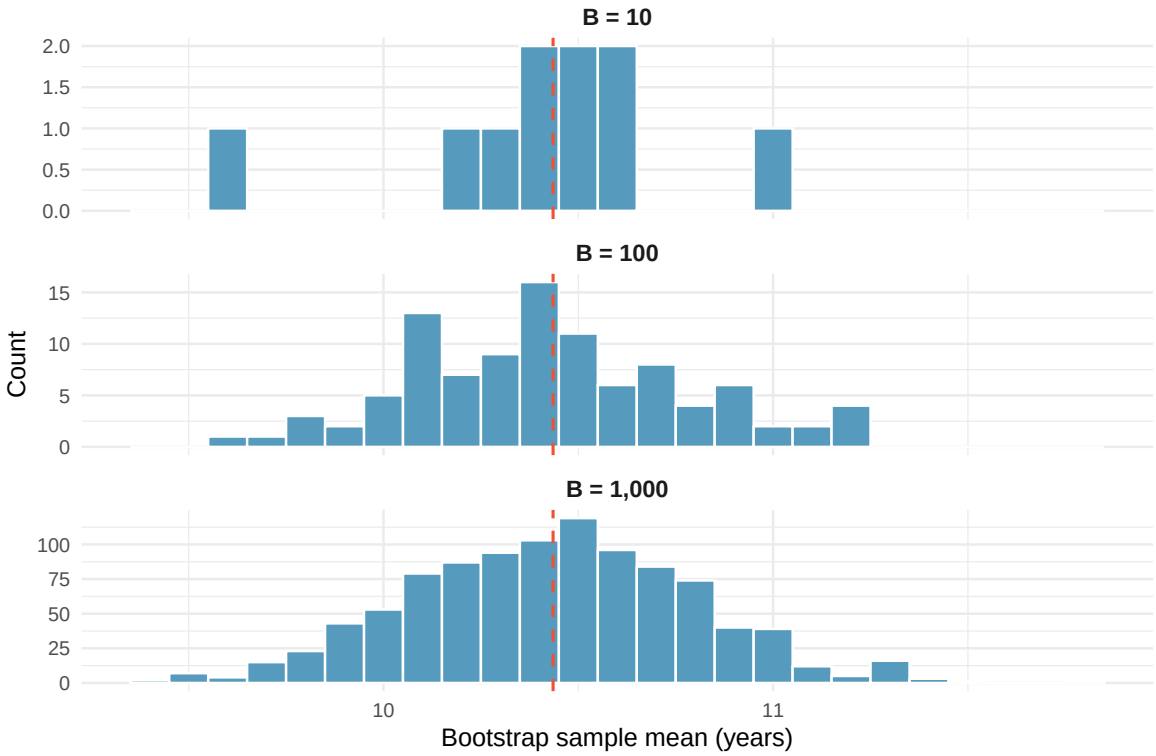


Figure 9.2: Building up the bootstrap distribution. With 10 resamples we see only a rough sketch; by 1,000 resamples the shape of the distribution is clear — and approximately bell-shaped (normal), even though the original data are skewed. Open in StatLens

Notice two things: (1) the bootstrap distribution is centered near the original sample mean (dashed red line), and (2) even though the penny ages are right-skewed, the bootstrap distribution of the *mean* is approximately bell-shaped (what statisticians call a **normal** shape — recall this term from the empirical rule in Section 4.7). This is the Central Limit Theorem at work — we'll explore it more in the next chapter.

Predict → Do → Explain. Open the Bootstrap Resampling Activity — the gated activity opens with a **Predict** step (commit to the bootstrap distribution's center and shape *before* the buildup starts), has you draw single resamples one at a time, watch the distribution build in an animation, then fill it out to 1,000 resamples and read off the 95% CI, and closes with an **Explain** step where you articulate why the bootstrap distribution centers where it does and why it looks bell-shaped even when the raw data are skewed. Your predictions and explanation save locally in the tool, so you can leave the activity and return to your work on the same device.

9.2.4 Step 4: Find the confidence interval

For a 95% confidence interval, we find the middle 95% of the bootstrap distribution by taking the **2.5th and 97.5th percentiles**.

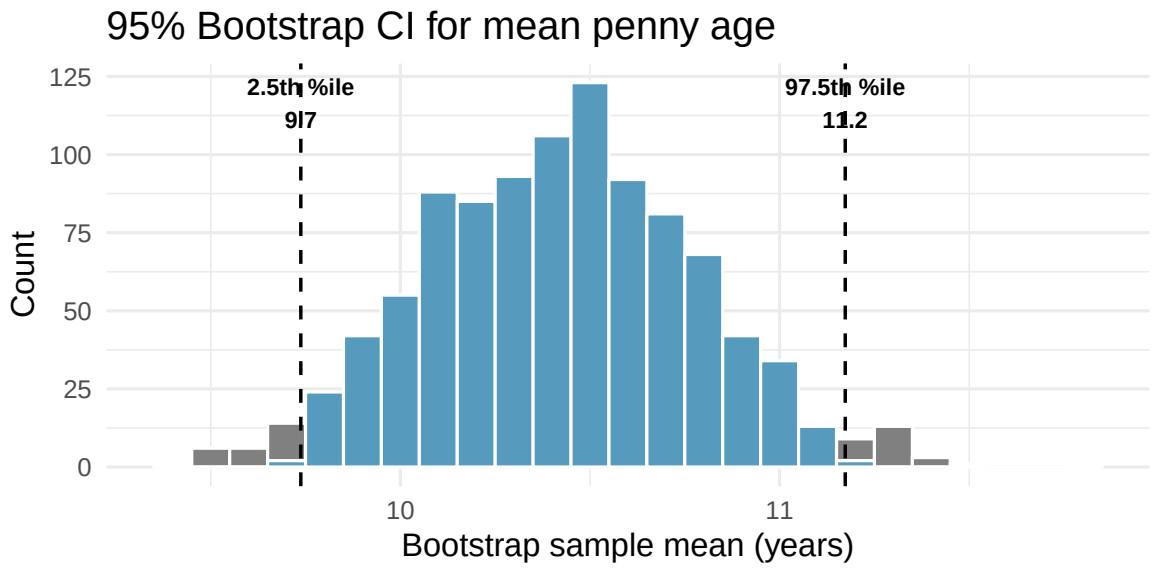


Figure 9.3: Bootstrap distribution of 1,000 sample means with the middle 95% shaded in blue. The 2.5th and 97.5th percentiles form the 95% confidence interval for the mean penny age. Open in StatLens

The 95% bootstrap confidence interval for the mean penny age is approximately $(9.7, 11.2)$ years.

9.3 Interpreting a confidence interval

A **confidence interval** is a range of plausible values for the population parameter. A **95% confidence interval** is constructed using a procedure that, when applied repeatedly to many different samples, captures the true parameter in approximately 95% of all intervals produced.

What "95% confident" means: If we repeated the entire study many times — collecting a new sample each time and computing a new 95% confidence interval — approximately 95% of those intervals would contain the true population parameter. Any single interval either contains the parameter or it doesn't; the 95% refers to the long-run success rate of the *procedure*.

To illustrate the coverage concept, imagine we could treat our 648 pennies as if they were the entire population (with true mean $\mu \approx 10.4$ years). We then draw 25 smaller random samples of 50 pennies each and build a bootstrap CI from each one. This lets us check whether each interval captures the "true" mean — something we can never do with real data, because we don't know the true parameter.

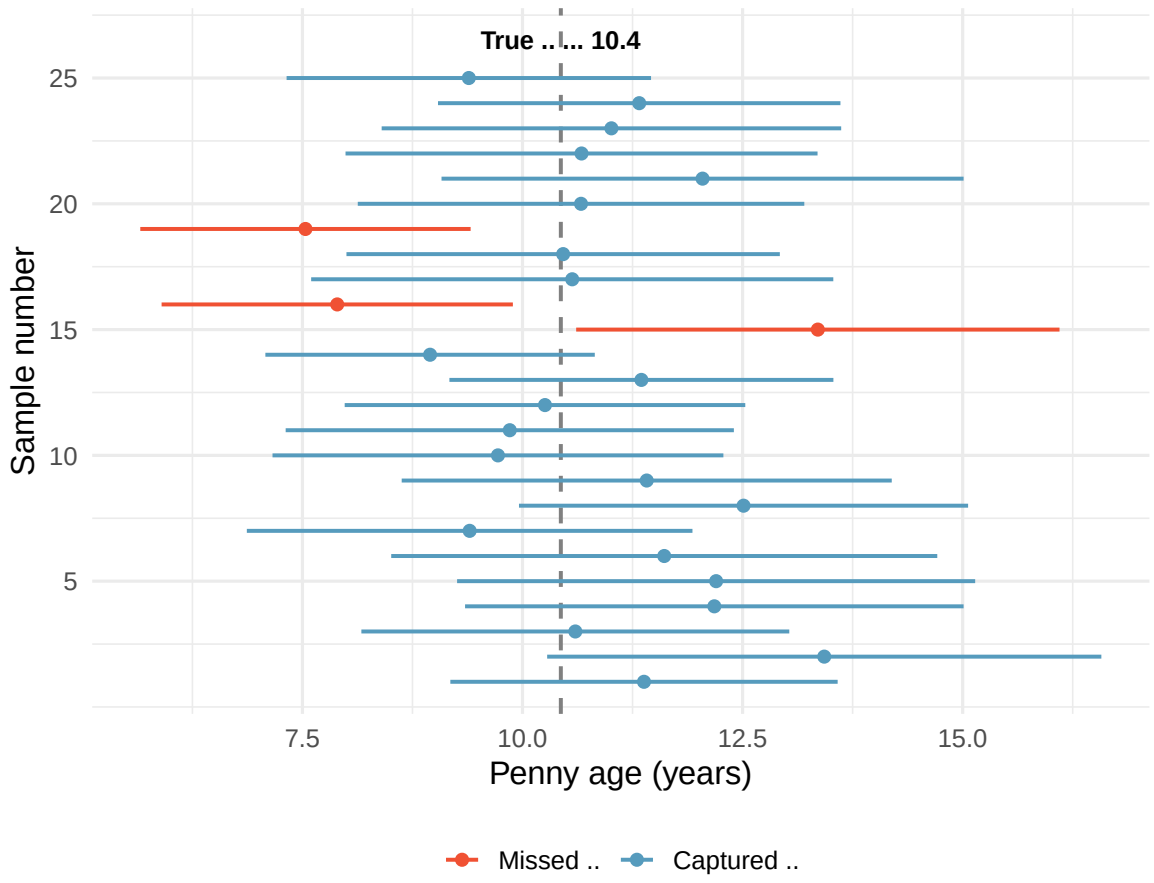


Figure 9.4: Twenty-five 95% bootstrap confidence intervals from 25 different random samples of pennies. The true population mean is shown as a vertical dashed line. In the long run about 95% of such intervals capture the true mean; here 22 of the 25 do, and the 3 shown in red miss it. Open in StatLens

Each horizontal line in Figure 9.4 represents a 95% CI from a different sample. Most intervals (blue) capture the true mean, but a few (red) miss it entirely. Over many samples, roughly 95% of the intervals succeed. This is the meaning of "95% confidence."

See it in action. Open the CI Coverage demo and keep drawing 95% confidence intervals from fresh samples of 50 pennies. Watch how most intervals (blue) capture the true mean while a few (red) miss — and how the running coverage rate settles near 95% as you accumulate more intervals. The moving image *is* the "95% confidence" concept in a way a static figure cannot be.

Common misinterpretation: "There is a 95% probability that the true mean is in this interval." This statement treats the population mean as random and the interval as fixed. In fact, the population mean is fixed (we just don't know it) and the interval is random (it depends on which sample we happened to collect).

A 95% bootstrap CI for the mean age of pennies in circulation is (9.7, 11.1). Which of the following interpretations is correct?

a. 95% of all pennies in circulation have ages between 9.7 and 11.1 years.
b. There is a 95% probability that the true mean penny age is between 9.7 and 11.1.
c. We are 95% confident that the true mean age of all pennies in circulation is between 9.7 and 11.1 years.

Show answer

Only (c) is correct. Statement (a) describes individual penny ages, not the mean. Statement (b) assigns a probability to the fixed parameter, which is incorrect. Statement (c) correctly describes confidence in the procedure.

9.4 The percentile method

The approach we've used is called the **percentile method** for bootstrap confidence intervals.

For a $(1 - \alpha) \times 100\%$ confidence interval using the percentile method:

- **Lower bound** = the $(\alpha/2) \times 100$th percentile of the bootstrap distribution
- **Upper bound** = the $(1 - \alpha/2) \times 100$th percentile of the bootstrap distribution

For a 95% CI ($\alpha = 0.05$): use the 2.5th and 97.5th percentiles. For a 90% CI ($\alpha = 0.10$): use the 5th and 95th percentiles. For a 99% CI ($\alpha = 0.01$): use the 0.5th and 99.5th percentiles.

If you increase the confidence level from 95% to 99%, what happens to the width of the confidence interval? Why?

Show answer

The interval gets wider. A 99% CI must capture the true parameter more often than a 95% CI, so it must cast a wider net. The 0.5th percentile is farther from the center than the 2.5th percentile.

9.5 The standard error method

The percentile method reads the confidence interval directly off the two tails of the bootstrap distribution. A second approach — the **standard error method** — instead summarizes the bootstrap distribution by its *spread*. It is the bridge to the formula-based confidence intervals we develop with the normal and t distributions in Unit 3.

The **bootstrap standard error** (SE) is the standard deviation of the bootstrap distribution. When the bootstrap distribution is roughly symmetric and bell-shaped, an approximate 95% confidence interval is

$$\text{point estimate} \pm 2 \times SE.$$

The multiplier 2 comes from the normal model: about 95% of a bell-shaped distribution falls within two standard deviations of its center. (The more precise multiplier is 1.96, which we begin using in Unit 3.)

For the penny ages, the observed sample mean is 10.4 years and the bootstrap standard error — the standard deviation of the 1,000 bootstrap means — is 0.36 years. The standard error 95% confidence interval is

$$10.4 \pm 2(0.36) = (9.7,\ 11.2) \text{ years}.$$

This is essentially the same interval we found with the percentile method — as we should expect, because the bootstrap distribution of the mean penny age is nearly symmetric.

Two methods, usually one answer. For a symmetric, bell-shaped bootstrap distribution, the percentile method and the standard error method give almost identical intervals. They can disagree when the bootstrap distribution is **skewed** — there the percentile method, which follows the actual

shape of the distribution, is the safer choice. The standard error method matters because it connects directly to the formula-based confidence intervals of Unit 3, where the standard error comes from a formula instead of from resampling.

9.6 Medical consultant case study

People providing an organ for donation sometimes seek the help of a special medical consultant. These consultants assist the patient in all aspects of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery.

One consultant tried to attract patients by noting the average complication rate for liver donor surgeries in the US is about 10%, but her clients have had only 3 complications in the 62 liver donor surgeries she has facilitated. She claims this is strong evidence that her work reduces complications.

We estimate the true complication rate p using the sample proportion: $\hat{p} = 3/62 = 0.048$ (4.8%).

Is it possible to assess the consultant's claim — that the *reduction* in complications is *due to her work* — using these data?

No. The claim is causal, but the data are observational. Patients who can afford a medical consultant may also afford better medical care overall, which could lead to a lower complication rate. We cannot establish causation here. But we *can* estimate the consultant's true complication rate using a confidence interval.

To build a bootstrap CI for a proportion, we treat the 62 patients as marbles in a bag: 3 orange (complication) and 59 blue (no complication). We draw 62 marbles with replacement, compute $\hat{p}^*_{boot}$, and repeat 10,000 times.

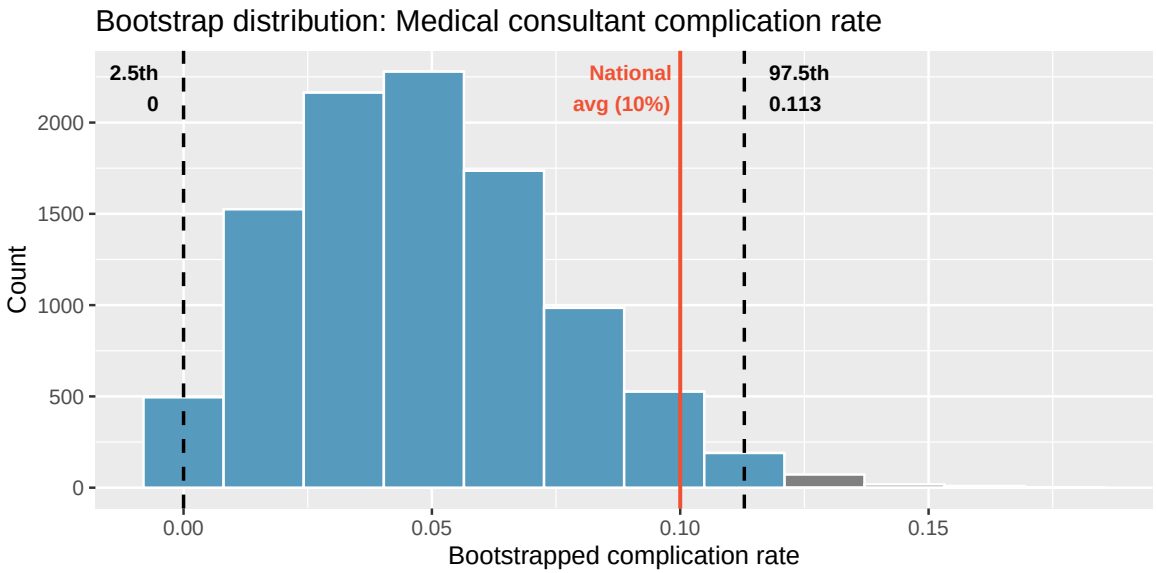


Figure 9.5: Bootstrap distribution of 10,000 resampled complication rates. The 95% CI extends from the 2.5th percentile to the 97.5th percentile. Because the interval includes 10%, we cannot conclude the consultant's rate is lower than the national average. Open in StatLens

The 95% bootstrap CI is approximately $(0, 0.113)$. The national average of 10% is right near the upper boundary of this interval — a borderline case. We certainly don't have strong evidence that the

consultant's rate is lower than the national average. The sample is small ($n = 62$) and the CI is wide, so complication rates anywhere from 0% up to about 11% are plausible.

Because there are only 62 patients and each is either "complication" or "no complication," the bootstrap proportions can only take values that are multiples of $1/62$ ($0, 1/62, 2/62, \ldots$). This **discreteness** means the CI boundary can only land on one of these values — it can't fall smoothly between them. You may notice the CI upper bound jump between $6/62 \approx 0.097$ and $7/62 \approx 0.113$ depending on how many resamples you generate. This is normal with small-sample proportion data and is one reason why borderline conclusions should be stated cautiously. It also means the interval can hold more than the nominal 95% of the bootstrap distribution — here about 99% of the bootstrapped proportions fall inside it, because the boundaries cannot fall between the discrete values. (That is a statement about this one bootstrap distribution, not about how often the method would capture the true rate across repeated studies.)

Predict → Do → Explain. Open the Medical Consultant Activity — the gated activity opens with a **Predict** step (commit to whether the 95% CI will include the national 10%), walks you through **bootstrapping the consultant's data at** $n = 62$, where the discreteness is impossible to miss — the interval's lower bound sits right at 0 — and closes with an **Explain** step where you put in your own words what "the CI includes 10%" does and does not let us conclude, and why small-n discreteness makes this borderline case harder than it would be at a larger sample. Your prediction and explanation save locally in the tool, so you can leave the activity and return to your work on the same device. (The guided practice below carries the larger-n comparison.)

Suppose the consultant had facilitated 620 surgeries (instead of 62) with the same complication rate of 4.8% (about 30 complications). How would the bootstrap CI change? Would the conclusion change?

Show answer

With 10 times as much data, the bootstrap distribution would be much narrower. The 95% CI would likely fall entirely below 10%, providing strong evidence that the consultant's rate is lower. Larger samples give more precise estimates.

9.7 Tappers and listeners case study

A Stanford University graduate student named Elizabeth Newton conducted an experiment using a simple game: one person (the "tapper") taps out a well-known song on a desk, and another person (the "listener") tries to guess the song.[1] About 50% of tappers expected the listener would guess correctly.

In Newton's study, only 3 out of 120 listeners ($\hat{p} = 3/120 = 0.025$) guessed the tune correctly. What is the true proportion of listeners who can guess the song?

[1] This case study is described in *Made to Stick* by Chip and Dan Heath.

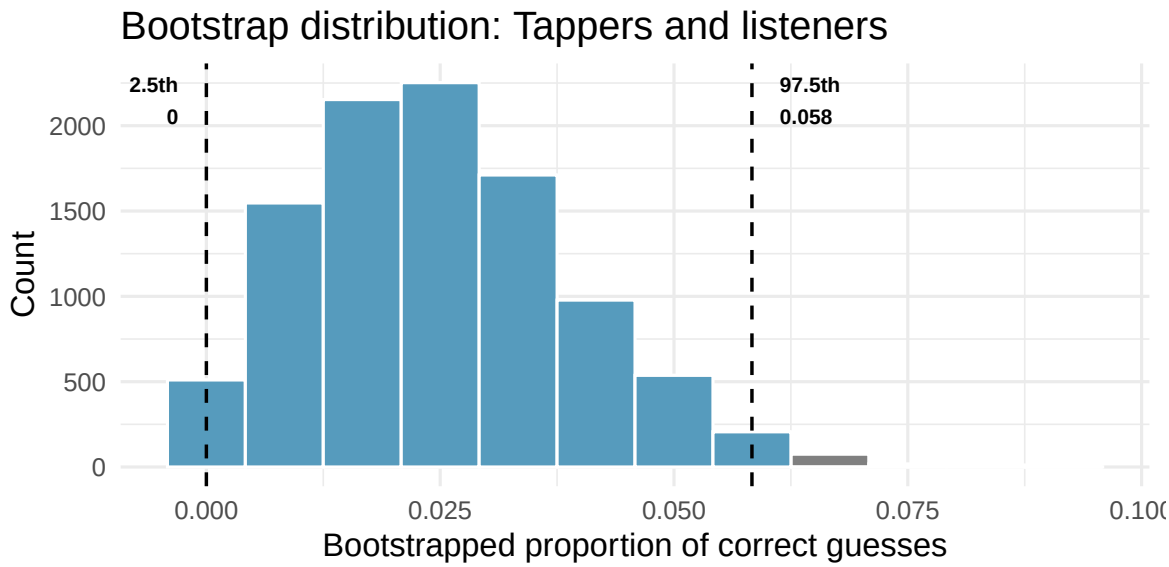


Figure 9.6: Bootstrap distribution for the proportion of listeners who guessed correctly. The 95% CI is far below the tappers' expected success rate of 50%, providing very convincing evidence against the tappers' expectation.

The 95% bootstrap CI is approximately $(0, 0.058)$. The tappers' expectation of 50% is nowhere near this interval — it provides **very convincing evidence** that listeners are far worse at guessing than tappers expect. This is an example of the "curse of knowledge": once you know something, it's hard to imagine not knowing it.

Do the data provide convincing evidence against the claim that 50% of listeners can guess the tapper's tune?

Show answer

Yes, very convincing evidence. The value 50% (0.50) is far outside the 95% CI, and even far from the largest bootstrapped proportion. The data strongly suggest that the true success rate is much lower than 50%.

9.8 What affects the width of a CI?

Three factors control how wide a confidence interval is:

1. **Sample size** (n): Larger samples → narrower intervals (more information about the population)
2. **Variability in the data**: More spread → wider intervals (less precision)
3. **Confidence level**: Higher confidence → wider intervals (more certainty requires a wider net)

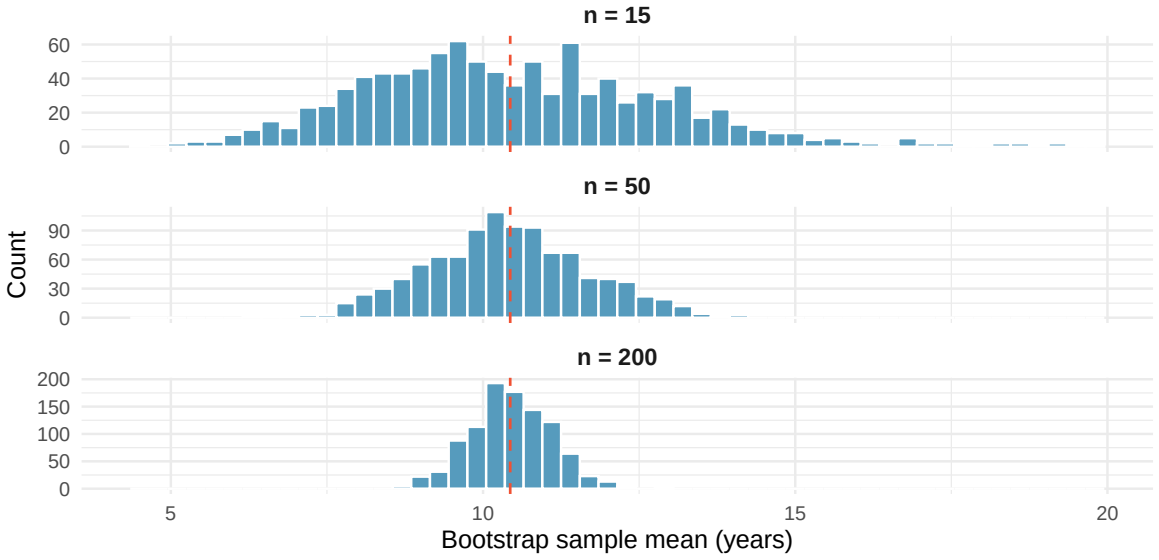


Figure 9.7: The effect of sample size on the bootstrap distribution. Larger samples produce narrower bootstrap distributions, leading to narrower confidence intervals. All three panels use the same population of penny ages. Open in StatLens

As sample size increases, the bootstrap distribution becomes tighter around the sample mean. This is why large studies produce more precise estimates than small ones.

Predict → Do → Explain. Open the Confidence Level Activity — the gated activity opens with a **Predict** step (commit to whether the interval gets wider or narrower as you move the confidence level from 90% to 95% to 99%), walks you through reading the 95%, 90%, and 99% intervals off the same bootstrap distribution so you can compare their widths directly, and closes with an **Explain** step where you articulate what "95% confident" means and why a wider net earns higher confidence. Your predictions and explanation save locally in the tool, so you can leave the activity and return to your work on the same device.

9.9 When does bootstrapping work?

Bootstrapping is remarkably flexible, but it does require some conditions to produce reliable intervals:

1. **Independence**: The observations in the sample should be independent of one another. This is usually satisfied when data come from a random sample or a randomized experiment.
2. **Representative sample**: The sample should be reasonably representative of the population. If the sample is biased (e.g., only volunteers), the bootstrap CI will reflect that bias.
3. **Sufficient sample size**: Very small samples (e.g., $n < 10$) may not contain enough information about the population's shape. With small n, the bootstrap distribution can be unreliable.

Bootstrapping does **not** require the population to follow a bell-shaped (normal) distribution. This is one of its greatest strengths compared to traditional formula-based methods, which often do require normality. The penny ages data are clearly right-skewed (Figure 9.1), yet the bootstrap procedure works perfectly well. The bootstrap lets the data "speak for themselves."

A researcher surveys 8 of her friends about their study hours and computes a bootstrap CI. What condition is most clearly violated?

Show answer

Two conditions are violated: (1) the sample is a convenience sample of friends, not a random sample, so it may not be representative of the population, and (2) $n = 8$ is quite small, making the bootstrap distribution potentially unreliable.

When neither bootstrap CI is enough. The two methods this chapter teaches — the **percentile** interval and the **standard-error** interval (statistic $\pm 2 \cdot SE_{\text{boot}}$) — are both starter tools, and they can go wrong in different ways. Be skeptical of *either* interval when:

- **Sample size is small** (below about 15–20). The bootstrap distribution can be lumpy and miss the shape of the true sampling distribution.
- **The parameter sits near a boundary** — a proportion near 0 or 1, a variance, an extreme quantile.
- **The statistic is unusual** — ratios, correlations, mins/maxes.

The two methods also have method-specific failure modes:

- The **standard-error method** assumes the bootstrap distribution is roughly symmetric and bell-shaped. When it isn't, the formula forces symmetry the data don't have and can push endpoints outside the parameter space (a negative bound for a proportion, for example).
- The **percentile method** respects skew but can under-cover the parameter on the "long-tail" side of a strongly skewed bootstrap distribution.

When any of these apply, professional practice reaches past both methods — most commonly to the **BCa** (bias-corrected and accelerated) interval, which corrects for both bias and skew in the bootstrap distribution. BCa is available in StatLens under **expert mode** and in R via `boot::boot.ci(..., type = "bca")`. Hesterberg's "What Teachers Should Know About the Bootstrap" (*The American Statistician*, 2015) is the readable reference for when each method applies.

For this course we stay with these two methods — they are the simplest defensible tools that keep the simulation logic in view. Recognize their limits: they are starters, not the final word.

9.10 Bootstrap CI for other statistics

The beauty of bootstrapping is its **generality**. The same resampling logic works for *any* statistic, not just the mean:

- **Median**: Resample → compute median → repeat → percentile CI
- **Proportion**: Resample → compute proportion → repeat → percentile CI
- **Standard deviation**: Resample → compute SD → repeat → percentile CI
- **Difference in means**: Resample from each group → compute difference → repeat → percentile CI

This is a powerful advantage. Many statistics (like the median or the difference in medians) don't have convenient formulas for their standard error. Bootstrapping handles them all the same way.

See it in action. Open the Bootstrap CI tool with penny ages and switch the statistic dropdown between the mean, median, standard deviation, Q1, and Q3. Same data, same resampling procedure — different summaries. Notice which statistics produce narrow, symmetric bootstrap distributions and which produce wide or skewed ones.

9.11 Chapter review

9.11.1 Summary

- The **bootstrap** constructs a confidence interval by resampling with replacement from the original sample.
- The **bootstrap distribution** approximates the sampling distribution of the statistic.
- The **percentile method** uses the middle $(1-\alpha) \times 100\%$ of the bootstrap distribution as the CI.
- The **standard error method** uses point estimate $\pm\, 2 \times SE$, where the SE is the standard deviation of the bootstrap distribution; it agrees with the percentile method for symmetric distributions and bridges to the formula-based intervals of Unit 3.
- A **95% confidence interval** means that if we repeated the study many times, about 95% of the resulting intervals would capture the true parameter.
- CI width depends on sample size, data variability, and confidence level.
- Bootstrap CIs work for any statistic — means, medians, proportions, differences, and more.
- Bootstrapping requires independent observations from a representative sample with sufficient sample size, but does **not** require the population to be bell-shaped (normal).

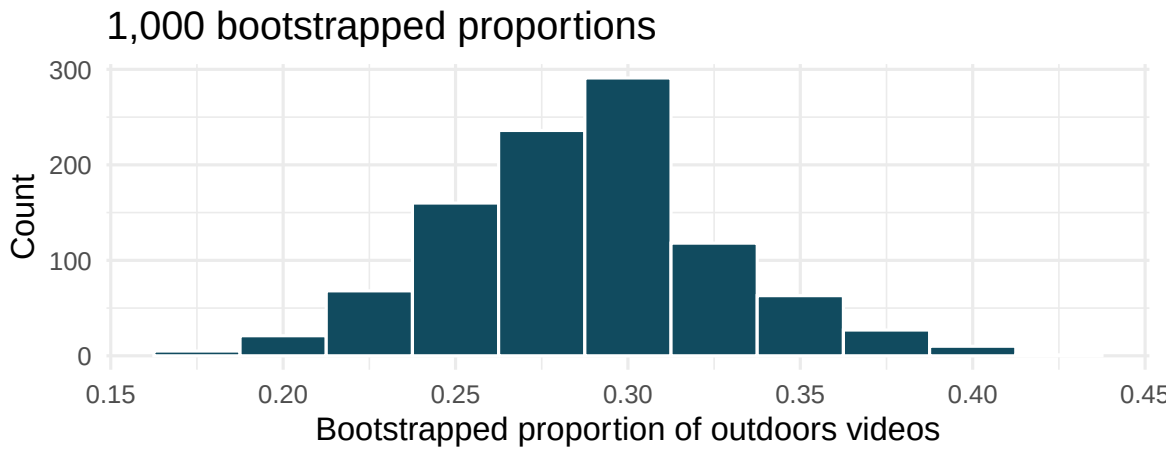
9.11.2 Key terms

- Bootstrap resample
- Bootstrap statistic
- Bootstrap distribution
- Confidence interval
- Confidence level
- Percentile method
- Standard error method
- Bootstrap standard error
- Point estimate
- Parameter
- Sampling with replacement

9.12 Exercises

The **StatLens Exercises** (1–14) use interactive simulation tools — open the provided links and explore. The **Practice Exercises** (15–22) use static figures for additional practice. Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Outside YouTube videos.** Let's say that you want to estimate the proportion of YouTube videos which take place outside (define "outside" to be if any part of the video takes place outdoors). You take a random sample of 128 YouTube videos and determine that 37 of them take place outside. You'd like to estimate the proportion of all YouTube videos which take place outside, so you decide to create a bootstrap interval from the original sample of 128 videos.



a. Describe in words the relevant statistic and parameter for this problem. If you know the numerical value for either one, provide it. If you don't know the numerical value, explain why the value is unknown.

b. What notation is used to describe, respectively, the statistic and the parameter?

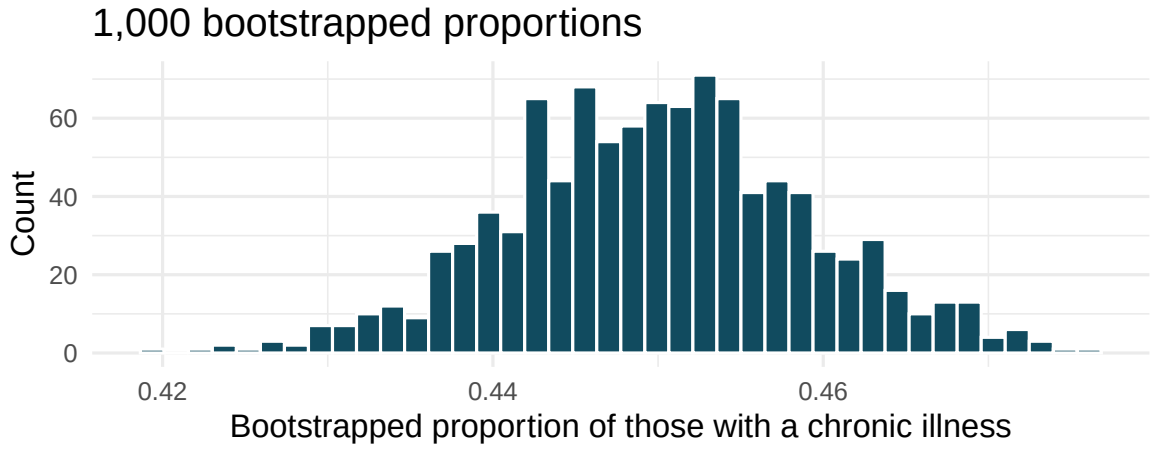
c. If using software to bootstrap the original dataset, what is the statistic calculated on each bootstrap sample?

d. When creating a bootstrap sampling distribution (histogram) of the bootstrapped sample proportions, where should the center of the histogram lie?

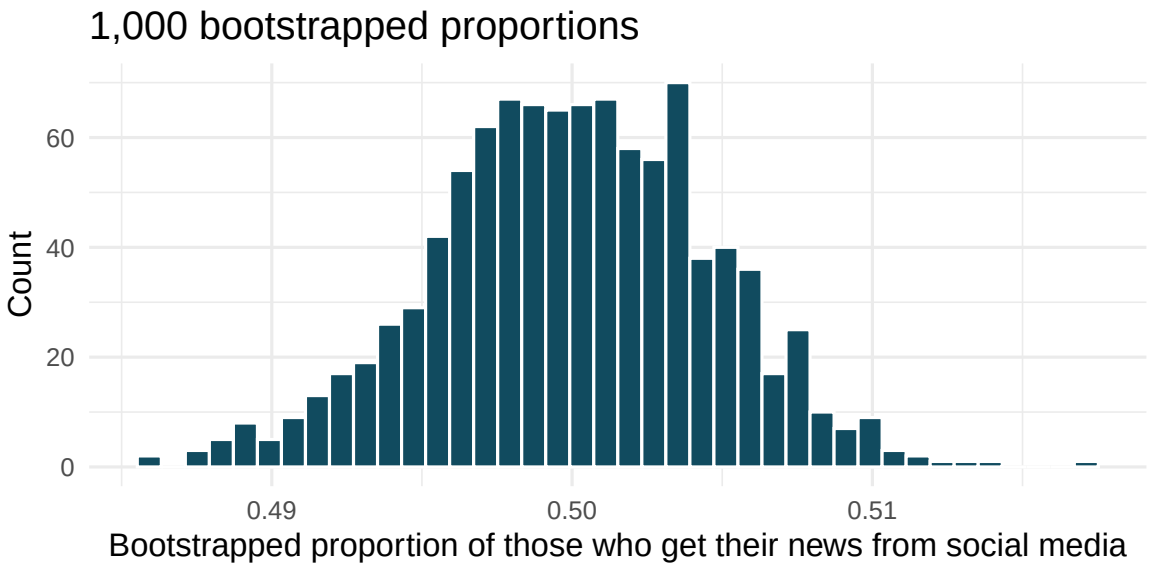
e. The histogram provides a bootstrap sampling distribution for the sample proportion (with 1000 bootstrap repetitions). Using the histogram, estimate a 90% confidence interval for the proportion of YouTube videos which take place outdoors.

f. Interpret the confidence interval in context of the data.

2. **Chronic illness.** In 2012 the Pew Research Foundation reported that "45% of US adults report that they live with one or more chronic conditions." However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study was based on a sample of 3014 adults. Below is a distribution of 1000 bootstrapped sample proportions from the Pew dataset. (Pew Research Center 2013) Using the distribution of 1,000 bootstrapped proportions, approximate a 92% confidence interval for the true proportion of US adults who live with one or more chronic conditions and interpret it.



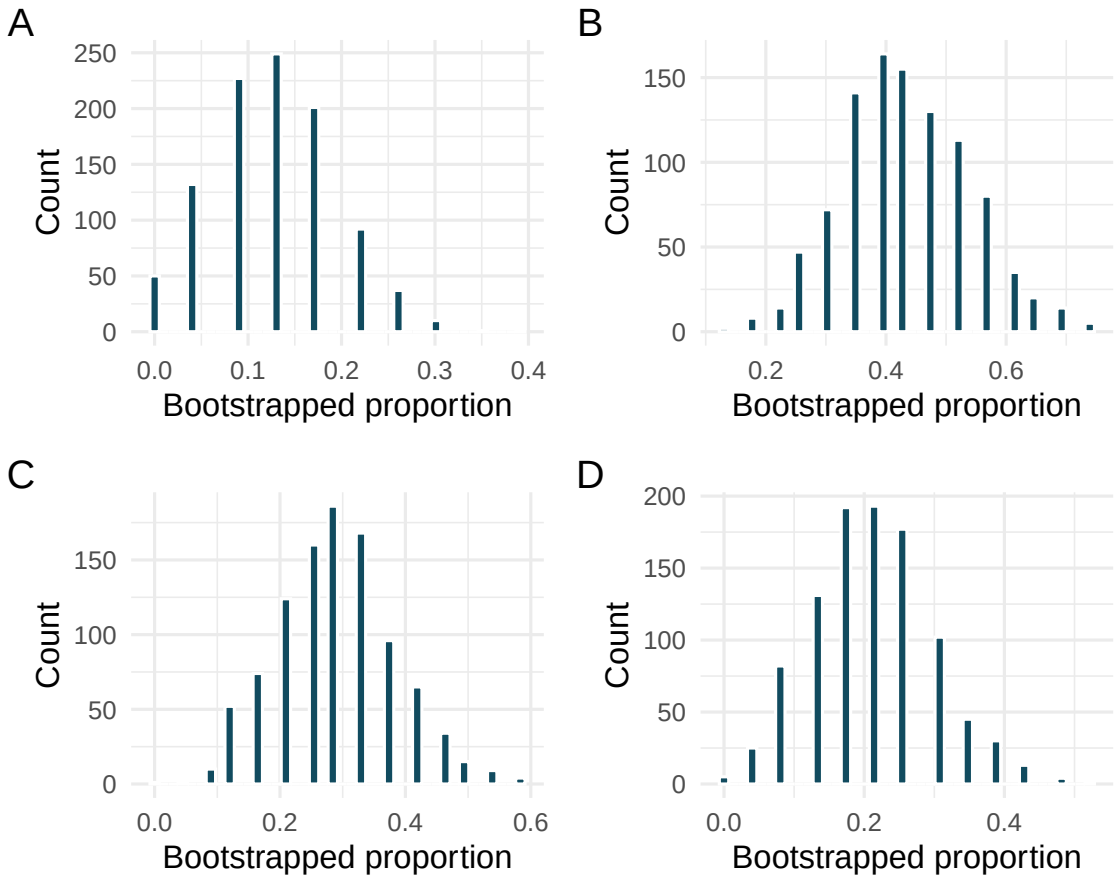
3. **Social media users and news, bootstrapping.** A poll conducted in 2022 found that 50% of U.S. adults get news from social media sometimes or often. However, the value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study was based on a sample of 12,147 adults. Below is a distribution of 1,000 bootstrapped sample proportions from the Pew dataset. (Pew Research Center 2022) Using the distribution of 1,000 bootstrapped proportions, approximate a 98% confidence interval for the true proportion of US adult social media users (in 2022) who get at least some of their news from Twitter. Interpret the interval in the context of the problem.



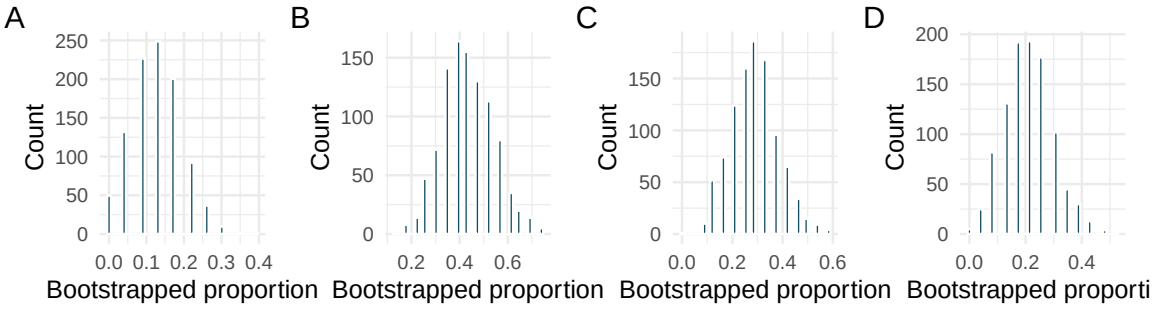
4. **Bootstrap distributions of $\hat{p}$, I.** Each of the following four distributions was created using a different dataset. Each dataset was based on $n = 23$ observations. The original datasets had the following proportions of successes:

$$\hat{p} = 0.13 \;\; \hat{p} = 0.22 \;\; \hat{p} = 0.30 \;\; \hat{p} = 0.43.$$

Match each histogram with the original data proportion of success.


A
Count
250
200
150
100
50
0
0.0
0.1
0.2
0.3
0.4
Bootstrapped proportion
B
Count
150
100
50
0
0.2
0.4
0.6
Bootstrapped proportion
C
Count
150
100
50
0
0.0
0.2
0.4
0.6
Bootstrapped proportion
D
Count
200
150
100
50
0
0.0
0.2
0.4
Bootstrapped proportion

5. **Bootstrap distributions of $\hat{p}$, II.** Each of the following four distributions was created using a different dataset. Each dataset was based on $n = 23$ observations.



Consider each of the following values for the true popluation p (proportion of success). Datasets A, B, C, D were bootstrapped 1000 times, with bootstrap proportions as given in the histograms provided. For each parameter value, list the datasets which could plausibly have come from that population. (Hint: there may be more than one dataset for each parameter value.)

a. $p = 0.05$

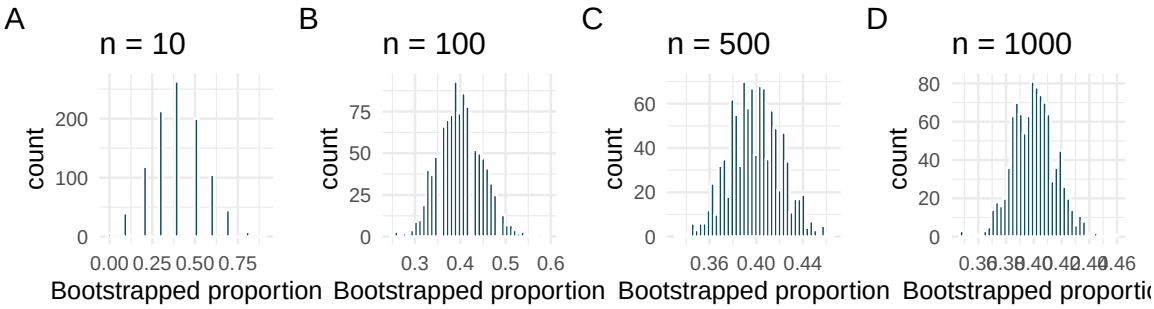
b. $p = 0.25$

c. $p = 0.45$

d. $p = 0.55$

e. $p = 0.75$

6. **Bootstrap distributions of $\hat{p}$, III.** Each of the following four distributions was created using a different dataset. Each dataset had the same proportion of successes ($\hat{p} = 0.4$) but a different sample size. The four datasets were given by $n = 10, 100, 500$, and 1000.



Consider each of the following values for the true popluation p (proportion of success). Datasets A, B, C, D were bootstrapped 1000 times, with bootstrap proportions as given in the histograms provided. For each parameter value, list the datasets which could plausibly have come from that population. (Hint: there may be more than one dataset for each parameter value.)

a. $p = 0.05$

b. $p = 0.25$

c. $p = 0.45$

d. $p = 0.55$

e. $p = 0.75$

7. **Cyberbullying rates.** Teens were surveyed about cyberbullying, and 54% to 64% reported experiencing cyberbullying (95% confidence interval). Answer the following questions based on this interval. (Pew Research Center 2018)

a. A newspaper claims that a majority of teens have experienced cyberbullying. Is this claim supported by the confidence interval? Explain your reasoning.

b. A researcher conjectured that 70% of teens have experienced cyberbullying. Is this claim supported by the confidence interval? Explain your reasoning.

c. Without actually calculating the interval, determine if the claim of the researcher from part (b) would be supported based on a 90% confidence interval?

8. **Waiting at an ER.** A 95% confidence interval for the mean waiting time at an emergency room (ER) of (128 minutes, 147 minutes). Answer the following questions based on this interval.

a. A local newspaper claims that the average waiting time at this ER exceeds 3 hours. Is this claim supported by the confidence interval? Explain your reasoning.

b. The Dean of Medicine at this hospital claims the average wait time is 2.2 hours. Is this claim supported by the confidence interval? Explain your reasoning.

c. Without actually calculating the interval, determine if the claim of the Dean from part (b) would be supported based on a 99% confidence interval?

9. **Percentile or standard-error method?** For each of the following bootstrap distributions, decide whether the **percentile method** and the **standard-error method** would give roughly the same 95% CI, or whether one is more trustworthy. Explain in one sentence.

a. The bootstrap distribution of the sample mean is roughly symmetric and bell-shaped, centered near the observed statistic.

b. The bootstrap distribution of the sample median is strongly right-skewed with a long upper tail.

c. The bootstrap distribution of a sample proportion $\hat{p}$ is bounded (values can't exceed 1) and pushed up against the upper boundary near $\hat{p} = 0.95$.

d. A study has $n = 12$; the bootstrap distribution looks lumpy and irregular but is roughly centered.

e. In one sentence, explain the underlying reason the two methods can disagree.

StatLens Exercises

These exercises focus on the parts of bootstrap inference a calculator *can't* do for you: **framing the question**, **picking the right resampling scheme**, **reading the bootstrap distribution**, and **interpreting** a confidence interval without falling into the standard traps. Let StatLens do the arithmetic.

1. **Frame the question.** A coffee shop wants to estimate the **median** wait time at its drive-through. The owner records a random sample of $n = 80$ wait times in seconds.
   a. What is the parameter of interest? Describe it in words and give a symbol.
   b. Why is a **bootstrap** confidence interval especially well-suited to the median compared to the mean?
   c. Should the resulting bootstrap CI be **percentile-based** or **standard-error-based**, and how would you choose between them?

2. **Choose the resampling scheme.** For each scenario, name the resampling — **resample with replacement from one sample** (one mean/median/SD/proportion), **resample within each group** (two means/proportions, independent groups), **resample paired-difference values** (paired data), or **the bootstrap doesn't apply here** (and why).
   a. Estimate the mean weight of a single bag of coffee from 30 weighed bags.
   b. Estimate the difference in mean delivery time between two couriers based on 25 deliveries each.
   c. Estimate the difference in the proportion of customers who return for a second visit after two different welcome promotions (random assignment).
   d. Estimate the standard deviation of completion times in a single sample of 60 jobs.
   e. Estimate the maximum age in a population from a random sample of 50 people.

3. **Read the bootstrap distribution.** Open the Bootstrap CI for One Proportion with the medical consultant data (62 patients, 3 complications, $\hat{p} \approx 0.048$). Generate 10000 resamples.
   a. Describe the **shape**, **center**, and **spread** of the bootstrap distribution. What feature of the data (small $\hat{p}$) is responsible for the shape?
   b. Read off the 95% percentile interval. Does it contain the national average of 0.10?
   c. If you instead built a **symmetric** 95% interval as $\hat{p} \pm 2 \cdot \text{SE}_{\text{boot}}$, what would go wrong? Sketch what that interval would look like next to the percentile interval.

4. **Run it and interpret it.** Open any one-variable numeric dataset in the Bootstrap CI for One Mean tool and generate 5000 resamples at 95% confidence.
   a. Report the 95% percentile interval StatLens gives you.
   b. Write a one-sentence conclusion **in context**, naming the population and the units.
   c. Which statement is the correct interpretation of the interval?
      i. There is a 95% probability that the true mean lies in this interval.
      ii. If the study were repeated many times, about 95% of the intervals built this way would contain the true mean.
      iii. 95% of the observations in our sample fall within this interval.
      iv. 95% of the bootstrap resamples have means in this interval.
   d. Change the confidence level to 99% and re-run. Does the interval get **wider** or **narrower**? Why?

5. **Simulation vs. analytic — where the bootstrap shines.** The normal-approximation formula and the bootstrap are two routes to the same confidence interval. Sometimes they nearly coincide; sometimes they don't.

First a well-behaved case. Open a moderately large, roughly symmetric numeric dataset in both the Bootstrap CI for One Mean and the One-Sample t tool.

a. Record the 95% interval from each. How close are they?

Now the troublesome case. Pick a *small*, *skewed* dataset (you can paste in your own).

b. Report the 95% intervals from each tool. Are they as close as in (a)? Where do they differ — in the **midpoint**, the **width**, or both?
c. Explain the pattern. Why do the bootstrap percentile interval and the t-interval agree on well-behaved data but diverge on small skewed data? Tie your explanation to (i) the symmetry of the t-interval and (ii) the bootstrap's ability to follow the shape of the sampling distribution.
d. A student says: "The t-interval and the bootstrap interval disagree, so one of them must be wrong." Why is that the wrong framing? Which interval should you *trust more* when they disagree, and why?

Chapter 10

Sampling Distributions via Simulation

In the randomization tests and bootstrap confidence intervals chapters, we used simulation to answer two fundamental questions of inference: *Is there an effect?* (randomization tests) and *How big is the parameter?* (bootstrap confidence intervals). Both methods relied on the computer to generate distributions — randomization distributions and bootstrap distributions — that helped us quantify uncertainty. In this chapter, we step back to see the bigger picture. We examine the shape, center, and spread of these simulated distributions and discover a remarkable pattern: under certain conditions, they all look approximately **normal** (bell-shaped). This observation, formalized as the **Central Limit Theorem**, is the bridge from the simulation-based methods of Unit 2 to the formula-based methods of Unit 3.

10.1 From simulation to theory

In the randomization tests and bootstrap confidence intervals chapters, we encountered several case studies:

- **Sex discrimination** (the randomization tests chapter): We shuffled promotion decisions and computed the difference in promotion rates 100 times.
- **Opportunity cost** (the randomization tests chapter): We shuffled buying decisions and computed the difference in proportions 1,000 times.
- **Medical consultant** (the bootstrap confidence intervals chapter): We bootstrapped the complication rate from 62 surgeries 10,000 times.
- **Tappers and listeners** (the bootstrap confidence intervals chapter): We bootstrapped the proportion of correct guesses from 120 listeners 10,000 times.

Each of these simulations produced a distribution of a sample statistic. While they differed in their settings, outcomes, and the simulation technique used, they all shared something in common: the general shape of the distribution.

Think back to the randomization and bootstrap distributions you have seen. What shape did most of them have? Were they skewed, uniform, or something else?

Show answer

In general, the distributions were approximately symmetric and bell-shaped. The two bootstrap distributions for small proportions — the medical consultant and the tappers and listeners — were the exceptions, with noticeable right skew, which makes sense because those proportions were very close to zero.

This is no coincidence. The bell shape keeps appearing because of a deep mathematical result that underlies most of statistics. Whether we shuffle labels, resample with replacement, or imagine taking repeated samples from a population, the distribution of the resulting statistic tends to look the same — symmetric, unimodal, and bell-shaped.

In the simulation chapters, we let the computer show us this pattern directly. Now, we formalize it.

10.2 The Central Limit Theorem

The **Central Limit Theorem** (**CLT**) states that when we take sufficiently large random samples from a population, the sampling distribution of many common statistics (such as $\hat{p}$ or $\bar{x}$) is approximately **normal** (bell-shaped), regardless of the shape of the underlying population distribution.

This is a stunning result. Even if the population distribution is skewed, bimodal, or has some other non-normal shape, the distribution of the *sample statistic* will be approximately normal — provided the sample is large enough.

Central Limit Theorem: the key ideas.

1. The sampling distribution of a sample statistic is approximately normal when the sample size is large enough.
2. The center of the sampling distribution is at the population parameter.
3. The spread of the sampling distribution (the standard error) decreases as sample size increases.

These three facts together mean: with a large enough sample, the sample statistic will be close to the population parameter, and we can use the normal distribution to describe *how* close.

10.2.1 Conditions for the CLT

The normal approximation does not apply in all situations. Two conditions must hold:

1. **Independent observations.** The observations in the sample must be independent of each other. This is typically guaranteed by random sampling from a population, or by random assignment in an experiment.
2. **Large enough sample.** The sample size cannot be too small. What counts as "large enough" depends on the context:
   - For proportions: we generally need at least 10 expected successes and 10 expected failures.
   - For means: the rule of thumb depends on how skewed the population is. For roughly symmetric populations, $n \geq 30$ is often sufficient. For highly skewed populations, larger samples may be needed.

When these conditions are not met, the normal approximation may be poor and simulation-based methods (randomization tests, bootstrap CIs) are the better choice. This is one of the great advantages of simulation: it works even when mathematical conditions are not satisfied.

10.2.2 Visualizing the CLT

Imagine a population that is distinctly non-normal — say, a right-skewed distribution like household income. If we take a single random sample of $n = 5$ people and compute the mean income, that mean might be quite far from the population mean μ. But if we do this thousands of times (each time taking a new sample of 5 and computing the mean), the distribution of those thousands of sample means will already start to look somewhat bell-shaped, even though the underlying population is skewed.

Now increase the sample size to $n = 30$. The distribution of sample means looks even more bell-shaped. At $n = 100$, it is nearly indistinguishable from a perfect normal curve.

This progression — from skewed population, to somewhat normal at small n, to very normal at larger n — is the Central Limit Theorem in action.

Sample size	Shape of sampling distribution
$n = 5$	Retains some skew from the population
$n = 30$	Approximately normal for most populations
$n = 100$	Very close to normal for virtually all populations

The speed at which the sampling distribution approaches normality depends on how non-normal the population is. Symmetric populations converge quickly; highly skewed populations take longer.

Predict → Do → Explain. Open the Sampling Distribution Lab to watch the CLT emerge from a strongly right-skewed population. The gated activity opens with a **Predict** step (commit to where the sample means will pile up when the population is skewed), walks you through drawing samples of $n = 10$ and then freezing that picture to compare it with $n = 100$ on a shared axis, and closes with an **Explain** step where you put the Central Limit Theorem in your own words — why the sampling distribution centers at μ and turns bell-shaped, and why a larger n shrinks the spread but does not move the center. Your predictions and explanation save locally in the tool, so you can return to your work on the same device.

10.3 Comparing randomization, bootstrap, and mathematical approaches

We now have three tools for statistical inference, each grounded in the same core ideas but using different mechanisms:

	Randomization	Bootstrap	Mathematical model
Question answered	Is there an effect? (hypothesis test)	How big is the parameter? (confidence interval)	Both
How variability is generated	Shuffle labels to simulate H_0	Resample with replacement from sample	Use formulas based on the normal distribution
What it models	Variability due to random assignment	Variability due to random sampling	Variability predicted by mathematical theory
Center of distribution	Null hypothesis value	Sample statistic	Depends on context (H_0 for tests, $\hat{p}$ for CIs)
Conditions required	Minimal	Minimal	Independence + large sample
Computational cost	Moderate (need many simulations)	Moderate (need many resamples)	Low (just plug into formula)

Key insight. All three methods answer the same fundamental question: *How much does the statistic vary?* They just measure that variability in different ways.

- **Randomization** asks: "How much would the statistic vary if the null hypothesis were true?"
- **Bootstrap** asks: "How much would the statistic vary across different samples from this population?"
- **Mathematical models** ask: "How much does theory predict the statistic would vary?"

When conditions are met, all three give similar answers. The mathematical approach is faster and requires no simulation, which is why it is widely used. But the simulation approaches are more flexible and work even when conditions for the mathematical model are not met.

10.3.1 Which method should I use?

Here is a practical guide:

1. **When conditions for the mathematical model are clearly met** (independent observations, large sample, not too skewed), the formula-based approach is convenient and widely accepted. This will be our primary tool in Unit 3.
2. **When conditions are questionable** (small sample, very skewed data, unusual statistic), simulation-based methods are safer. The bootstrap and randomization test make fewer assumptions.
3. **When you want to build intuition**, simulation is invaluable. Seeing 10,000 shuffles or resamples gives you a concrete feel for what "random variability" looks like — something a formula alone cannot provide.
4. **In practice**, many statisticians use simulation to check their formula-based results. If the two approaches agree, confidence in the conclusion is strengthened.

10.4 When is the normal approximation good enough?

The mathematical approach to inference (which we will develop in Unit 3) replaces simulation with the normal distribution. Instead of generating thousands of shuffles or resamples, we use a formula to compute the standard error and then rely on the bell curve to find p-values or construct confidence intervals.

But when can we trust this shortcut?

10.4.1 The success-failure condition for proportions

For a sample proportion $\hat{p}$, the sampling distribution is approximately normal when:

$$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10$$

where n is the sample size and p is the proportion of interest. In practice, since p is unknown, we use $\hat{p}$ as a stand-in:

$$n\hat{p} \geq 10 \quad \text{and} \quad n(1-\hat{p}) \geq 10$$

This is called the **success-failure condition**.

In the medical consultant example, $n = 62$ and $\hat{p} = 3/62 = 0.048$. Check the success-failure condition.

- Expected successes: $n\hat{p} = 62 \times 0.048 = 3$ (less than 10)
- Expected failures: $n(1-\hat{p}) = 62 \times 0.952 = 59$ (at least 10)

The success-failure condition is **not met** because there are fewer than 10 expected successes. The normal approximation would be unreliable here. This is exactly the kind of situation where the bootstrap (as used in the bootstrap confidence intervals chapter) is the better approach.

In the tappers and listeners study, $n = 120$ and $\hat{p} = 3/120 = 0.025$. Check the success-failure condition. Should we use the normal approximation?

Show answer

Expected successes: $120 \times 0.025 = 3$ (less than 10). Expected failures: $120 \times 0.975 = 117$ (at least 10). The condition is not met because there are too few successes. The normal approximation would not be appropriate. Bootstrap methods are preferred.

A poll of $n = 1{,}000$ voters finds that $\hat{p} = 0.52$ support a policy. Check the success-failure condition.

- Expected successes: $n\hat{p} = 1{,}000 \times 0.52 = 520$ (well above 10)
- Expected failures: $n(1 - \hat{p}) = 1{,}000 \times 0.48 = 480$ (well above 10)

The success-failure condition is comfortably met. The normal approximation is reliable here.

10.4.2 The sample size condition for means

For a sample mean $\bar{x}$, the CLT tells us the sampling distribution is approximately normal when:

- The sample size is "large enough," and
- The observations are independent.

A common guideline is $n \geq 30$, but this is only a rough rule of thumb. If the population distribution is nearly symmetric, the normal approximation can work well even for smaller samples. If the population is highly skewed or has extreme outliers, a larger sample may be needed.

Population shape	Minimum n for normal approximation
Symmetric, no outliers	$n \geq 15$ or even smaller
Slightly skewed	$n \geq 30$
Highly skewed or heavy-tailed	$n \geq 60$ or more

When in doubt, use simulation (bootstrap) to check whether the normal approximation is reasonable.

10.4.3 Simulation as a check on theory

One of the most powerful uses of simulation is as a **reality check** on mathematical results. Here is the workflow:

1. Compute a confidence interval or p-value using the formula-based (normal) approach.
2. Compute the same quantity using simulation (bootstrap CI or randomization test).
3. Compare. If the two results are similar, you can be confident in the mathematical approach. If they differ substantially, the conditions for the normal approximation may not be met, and you should rely on the simulation result.

This "trust but verify" approach is standard practice among professional statisticians and data scientists.

10.5 The normal distribution: a preview

Since the normal distribution will be our primary tool in Unit 3, let us briefly introduce its key features here.

10.5.1 The 68-95-99.7 rule

The normal distribution has a very useful property: the percentage of observations falling within 1, 2, and 3 standard deviations of the mean is always approximately the same.

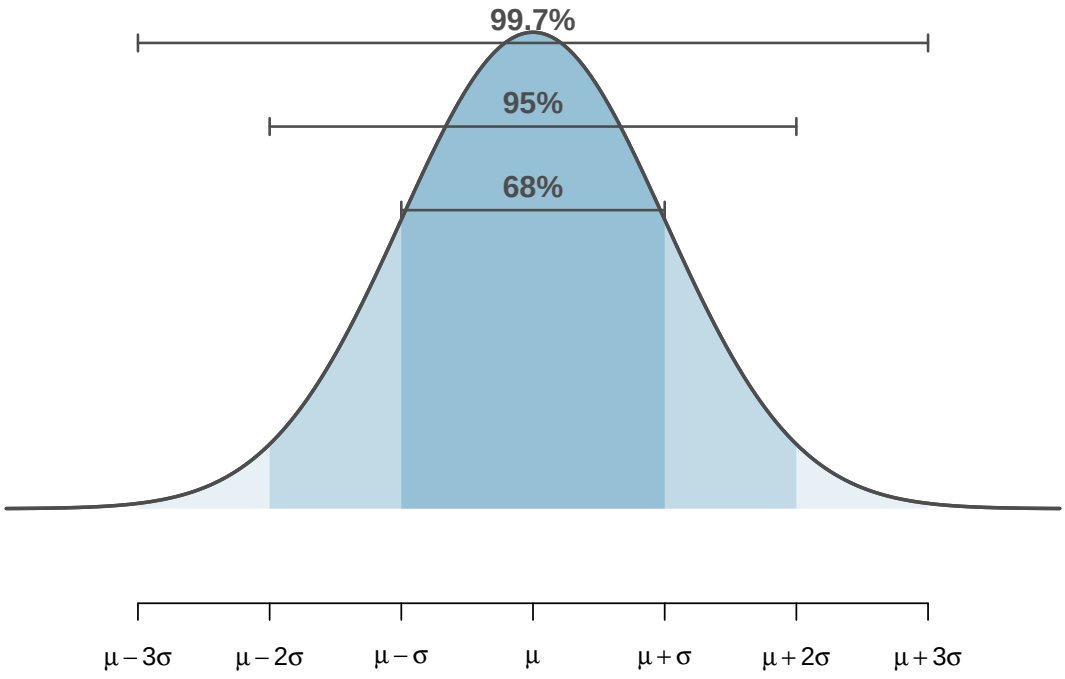


Figure 10.1: The 68-95-99.7 rule for normal distributions. Approximately 68% of values fall within 1 standard deviation of the mean, 95% within 2, and 99.7% within 3. ↗ Try this live — drag the cutlines to see the 68-95-99.7 boundaries live.

The 68-95-99.7 rule.

For any normal distribution:

- About **68%** of values fall within 1 standard deviation of the mean: $\mu \pm \sigma$
- About **95%** of values fall within 2 standard deviations of the mean: $\mu \pm 2\sigma$
- About **99.7%** of values fall within 3 standard deviations of the mean: $\mu \pm 3\sigma$

This rule is the reason behind the **quick approximation we already met in the bootstrap confidence intervals chapter**: a 95% confidence interval $\approx$ statistic $\pm\, 2 \times SE$. We used it there; now we can see *why* it works — when the bootstrap (or sampling) distribution is roughly normal, its middle 95% lies within about 2 standard deviations, which for a statistic means about 2 standard errors:

$$\text{95\% CI} \approx \text{statistic} \pm 2 \times SE$$

A poll of $n = 1{,}000$ voters finds $\hat{p} = 0.52$ with a standard error of $SE = 0.016$. Use the quick approximation to construct a 95% confidence interval.

$$\hat{p} \pm 2 \times SE = 0.52 \pm 2(0.016) = 0.52 \pm 0.032 = (0.488, 0.552)$$

We are 95% confident that the true proportion of voters who support the policy is between 48.8% and 55.2%.

Why is it "approximately 2" standard errors rather than exactly 2? What is the more precise multiplier for a 95% confidence interval based on the normal distribution?

Show answer

The more precise multiplier is 1.96, not 2. The value 1.96 is the Z-score that puts exactly 2.5% in each tail of the standard normal distribution, leaving 95% in the middle. The approximation of 2 is close enough for quick mental calculations.

10.5.2 Z scores

A **Z score** measures how many standard deviations an observation is above or below the mean:

$$Z = \frac{x - \mu}{\sigma}$$

In the context of sampling distributions, we can compute a Z score for an observed statistic:

$$Z = \frac{\text{statistic} - \text{parameter}}{SE}$$

This Z score tells us how unusual the observed statistic is, measured in standard-error units. In Unit 3, we will use Z scores (and their cousins, t-scores) extensively to compute p-values and confidence intervals without simulation.

In the sex discrimination study, the observed difference in promotion rates was 0.292. Suppose the standard error of the difference (computed under the null hypothesis) is 0.12. Compute the Z score.

Under H_0, the parameter (the true difference) is 0. So:

$$Z = \frac{0.292 - 0}{0.12} = 2.43$$

The observed difference is 2.43 standard errors above the null value of 0. Since values beyond 2 standard deviations are unusual for a normal distribution (the 95% rule), this is strong evidence against the null hypothesis — consistent with our earlier simulation result.

10.6 Recap: connecting the three approaches

Types of distributions (a reference guide).

You have now encountered several types of distributions in this course. Understanding the differences among them is essential:

- A **data distribution** (or sample distribution) describes the shape, center, and variability of the observed data.
- A **population distribution** describes the entire population — almost always unknown.
- A **sampling distribution** describes how a *sample statistic* varies from sample to sample. The CLT tells us this is approximately normal under certain conditions.
- A **randomization distribution** describes how a sample statistic varies under random shuffling of the treatment variable. It is centered at the null hypothesis value.
- A **bootstrap distribution** describes how a statistic varies under resampling from the observed data. It is centered at the observed statistic.

10.6.1 The big picture of Unit 2

Unit 2 has built your intuition for statistical inference using simulation:

Chapter	Method	Question	How variability is generated
6	Sampling variability	How much do statistics vary?	Conceptual + simulation
7	Randomization test	Is there an effect?	Shuffle labels
8	Decision errors	How can a test go wrong?	Simulated Type I / Type II rates
9	Bootstrap CI	How big is the parameter?	Resample with replacement
10	CLT and normal model	Can we skip the simulation?	Mathematical formula

The key takeaway: **the simulation methods and the mathematical methods are answering the same questions**. The CLT tells us *when* the mathematical shortcuts are valid. When they are, we can replace thousands of computer simulations with a single formula.

In Unit 3, we will develop these formula-based methods for specific settings: one proportion, one mean, two proportions, two means, and more. But you now have the conceptual foundation to understand *why* those formulas work: they are mathematical descriptions of the same sampling variability that you observed through simulation.

10.7 Chapter review

10.7.1 Summary

- Randomization distributions, bootstrap distributions, and theoretical sampling distributions all tend to be approximately **bell-shaped** (normal) under common conditions. This is the **Central Limit Theorem** (CLT).
- The CLT states that for sufficiently large samples with independent observations, the sampling distribution of many common statistics is approximately normal, regardless of the shape of the population.
- The **success-failure condition** ($n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$) is the practical check for whether the normal approximation is appropriate for proportions.
- The **68-95-99.7 rule** describes the area under a normal curve within 1, 2, and 3 standard deviations of the mean.
- A **Z score** measures how many standard deviations (or standard errors) an observation or statistic falls from the mean: $Z = (x - \mu)/\sigma$.

- The three inference approaches — randomization, bootstrap, and mathematical models — answer the same fundamental questions. When conditions are met, they produce similar results. Mathematical models are faster; simulation methods are more flexible.
- This chapter bridges Unit 2 (simulation-based inference) and Unit 3 (formula-based inference). The CLT is the theoretical justification for the formulas we will use in Unit 3.

10.7.2 Key terms

Term	Definition
Central Limit Theorem (CLT)	The sampling distribution of a statistic is approximately normal for large samples
Normal distribution	A symmetric, bell-shaped distribution described by its mean and standard deviation
Success-failure condition	For proportions: need $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$
68-95-99.7 rule	About 68%, 95%, 99.7% of normal values fall within 1, 2, 3 SDs of the mean
Z score	Number of standard deviations an observation falls from the mean: $Z = (x - \mu)/\sigma$
Null distribution	The sampling distribution of the sample statistic under H_0
Sampling distribution	Distribution of a statistic over all possible samples of a given size

10.8 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Repeated water samples.** A nonprofit wants to understand the fraction of households that have elevated levels of lead in their drinking water. They expect at least 5% of homes will have elevated levels of lead, but not more than about 30%. They randomly sample 800 homes and work with the owners to retrieve water samples, and they compute the fraction of these homes with elevated lead levels. They repeat this 1,000 times and build a distribution of sample proportions.

a. What is this distribution called?

b. Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

c. What is the name of the variability of this distribution.

d. Suppose the researchers' budget is reduced, and they are only able to collect 250 observations per sample, but they can still collect 1,000 samples. They build a new distribution of sample proportions. How will the variability of this new distribution compare to the variability of the distribution when each sample contained 800 observations?

2. **Repeated student samples.** Of all freshman at a large college, 16% made the dean's list in the current year. As part of a class project, students randomly sample 40 students and check if those students made the list. They repeat this 1,000 times and build a distribution of sample proportions.

a. What is this distribution called?

b. Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

c. What is the name of the variability of this distribution?

d. Suppose the students decide to sample again, this time collecting 90 students per sample, and they again collect 1,000 samples. They build a new distribution of sample proportions. How will the variability of this new distribution compare to the variability of the distribution when each sample contained 40 observations?

3. **Stating the Central Limit Theorem.** The Central Limit Theorem (CLT) is one of the most important results in statistics. It says that for a random sample of independent observations $X_1, X_2, \ldots, X_n$ drawn from a population with mean μ and finite standard deviation σ:

- The sampling distribution of the sample mean $\bar{X}$ has **mean** μ,
- The sampling distribution of $\bar{X}$ has **standard deviation** $\sigma/\sqrt{n}$,
- As n grows, the **shape** of the sampling distribution of $\bar{X}$ becomes approximately **normal**, **regardless of the shape of the population**.

a. A population of household incomes is strongly right-skewed. Which of the following are true according to the CLT? (Select all that apply — no justification needed.)

   i. As n grows, the sampling distribution of $\bar{X}$ becomes approximately normal.
   ii. As n grows, the population distribution becomes approximately normal.
   iii. As n grows, the standard error $\sigma/\sqrt{n}$ shrinks.
   iv. As n grows, the population mean μ changes.

b. A survey draws a random sample of $n = 400$ students; call the sample mean GPA $\bar{X}$. The population of student GPAs has mean $\mu = 3.15$ and standard deviation $\sigma = 0.60$. What are the approximate mean and standard error of the sampling distribution of $\bar{X}$?

c. A common misstatement of the CLT is: *"When n is large, the sample data will be approximately normal."* Explain in one sentence why this is wrong.

d. Roughly how large does n need to be for the CLT to give a good approximation? Discuss briefly.

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Predict, then check the CLT.** Open the Sampling Distribution Lab and pick a **right-skewed** population.
   a. Before drawing anything: at $n = 5$, do you expect the sampling distribution of $\bar{x}$ to (i) match the shape of the population, (ii) look symmetric and bell-shaped, or (iii) look somewhere in between?
   b. Draw 5000 sample means at $n = 5$. What shape do you actually see?
   c. Repeat at $n = 30$ and $n = 100$. At what point does the sampling distribution look "approximately normal" to your eye?
   d. Now switch the population to **bimodal**. At what n does the sampling distribution of $\bar{x}$ start to look normal, compared to the right-skewed case?

2. **More samples vs. larger n — which knob does what?** Stay in the Sampling Distribution Lab with a right-skewed population at $n = 10$.
   a. Draw 100 samples and note the *spread* (the displayed SD of the sample means). Then draw 5000 more (without changing n). Did the spread shrink?
   b. Now reset and start over at $n = 40$. Draw 5000 samples. How does the spread compare to your $n = 10$ run?
   c. Predict, without computing: if you go from $n = 10$ to $n = 40$, by what factor should the standard error of $\bar{x}$ shrink? Check against what the lab reports for $\sigma/\sqrt{n}$.
   d. A student says: "If I just draw enough samples, the standard error will eventually drop to zero." Why is this wrong? What *does* drop to (essentially) zero with more samples?

3. **Sampling distribution of $\hat{p}$.** Switch the Sampling Distribution Lab to **proportion mode**. Set the true population proportion to $p = 0.20$ and the sample size to $n = 20$.
   a. Predict the **center** and **spread** of the sampling distribution of $\hat{p}$ (use $\sqrt{p(1-p)/n}$ for the spread).
   b. Draw 5000 samples. Compare your predicted center and spread to the observed values. How close are they?
   c. Look at the **shape** of the sampling distribution. Is it symmetric or skewed? Why?
   d. Increase n to 200 (keep $p = 0.20$). Does the shape change? Use the success–failure rule ($np \geq 10$, $n(1-p) \geq 10$) to predict what should happen at $n = 200$ vs. $n = 20$.

4. **Reading sampling distributions in context.** A factory produces light bulbs with mean lifetime $\mu = 1000$ hours and standard deviation $\sigma = 200$ hours. A quality engineer plans to take a random sample of $n = 50$ bulbs and compute the sample mean lifetime $\bar{x}$.
   a. What is the **sampling distribution** of $\bar{x}$? Give its mean, standard error, and shape (and justify the shape).
   b. What is the probability the sample mean is *less than* 950 hours? Use a normal distribution tool (pre-loaded with $\mu = 1000$ and $\sigma = 200/\sqrt{50} \approx 28.28$).
   c. Suppose the engineer instead samples $n = 200$ bulbs. Without re-running the tool, will the probability in (b) be smaller or larger? Why?
   d. An executive says: "The sample mean has the same distribution as a single bulb's lifetime." Why is this wrong, and what's the impact on planning?

5. **The CLT's escape clauses — when do you *not* get normality?** The CLT requires (i) an i.i.d. sample and (ii) a *finite* variance.
   a. Suppose 50 students at one cafeteria table sample sizes their lunches. Why is this **not** an i.i.d. sample for the population of students? What might go wrong if you computed a

CLT-based interval for the mean lunch size?

b. The **Cauchy distribution** has no defined mean or variance — it's a heavy-tailed distribution. What would you expect to see if you drew sample means from a Cauchy population at $n = 30$, $n = 100$, $n = 1000$? (You can't simulate this in the lab, but reason about it.)

c. In practice, why is the violation in (a) (clustering / non-independence) much more common in real data than the violation in (b) (no finite variance)?

d. Suppose you have $n = 5000$ but the data come from social-media engagement (one big viral post can outweigh 4999 normal ones). Is the CLT comforting here? Why or why not?

Simulate — Unit Review

Unit 2 built **simulation-based inference** from the ground up. This review pulls the five chapters together: one idea — *use the computer to imitate chance* — applied two ways, to **test** a claim and to **estimate** a quantity, wrapped in a framework for the **errors** any decision can make. Nothing here is new; it is a map of what you should now be able to do, and how the pieces fit.

The big picture

Every method in this part rests on **sampling variability**: a statistic computed from a sample is just one draw from a distribution of values it *could* have taken. Simulation lets you build that distribution by hand instead of assuming a formula for it.

You used that one idea in two directions:

- **To test a claim** — a **randomization test**. Imitate a world in which the null hypothesis is true (by *shuffling* the group labels), and see how surprising your real result would be in that world. The answer is a **p-value**.
- **To estimate a quantity with uncertainty** — a **bootstrap confidence interval**. Imitate drawing new samples (by *resampling your one sample with replacement*), and watch how much the estimate bounces around. The answer is a **confidence interval**.

Decision errors gave you the language for what a test can get wrong, and **sampling distributions** showed that these simulated distributions have predictable shapes — the bridge to the formula-based methods of Unit 3.

Two engines, one idea: randomization vs. bootstrap

The two simulation methods are easy to confuse because both "shuffle the data and recompute" thousands of times. They answer *different questions* and manipulate the data in *different ways*:

	Randomization test	**Bootstrap interval**
Question it answers	*Is there an effect?* (test)	*How big is it?* (estimate)
You start by assuming...	the null hypothesis is **true**	nothing — you trust your sample
What you do to the data	**shuffle / relabel** (sample *without* replacement)	**resample *with* replacement**, same n
The simulated distribution is...	the **null distribution** of the statistic	the **sampling distribution** of the estimate
What you read off	a **p-value** (how extreme is the observed statistic?)	a **confidence interval** (percentile or SE method)

	Randomization test	Bootstrap interval
Conclusion sounds like	"discernible / no discernible evidence of an effect"	"we're 95% confident the value is between ___ and ___"

The tell: **a test shuffles labels to break the link** between groups; **a bootstrap resamples with replacement to imitate re-sampling the population.** If the question has the word *significant*/*discernible* or *evidence*, you are testing. If it asks *how much* or *what is the value*, you are estimating.

Which simulation do I run?

1. **What is the question?**
   - "Is there a difference / an effect / evidence against a claim?" → **randomization test → p-value.**
   - "What is the value, and how precise is my estimate?" → **bootstrap → confidence interval.**
2. **How many groups?** One group vs. a fixed claim, or two groups compared.
3. **What kind of variable?** A **proportion** (categorical outcome) or a **mean** (numerical outcome).

Those three answers name the exact procedure (e.g. *two groups + proportion + "is there an effect" → randomization test for a difference in proportions*).

Key ideas to carry forward

Sampling distribution · **null distribution** · **p-value** · **discernibility level** (α) · **Type I error** (rejecting a true null) · **Type II error** (missing a real effect) · **power** · **confidence level** · **bootstrap** · **percentile method** · **standard-error method**.

Common pitfalls

- **The p-value is not the probability the null is true.** It is the probability of data this extreme *if* the null were true.
- **A confidence interval is a statement about the procedure, not this one interval.** "95% confident" means 95% of intervals built this way would capture the truth — not a 95% probability for these particular endpoints.
- **"Fail to reject" is not "prove the null."** Absence of discernible evidence is not evidence of absence.
- **Randomization shuffles *without* replacement; the bootstrap resamples *with* replacement.** Mixing these up is the most common Unit 2 error.
- **α is chosen before you see the data.** It *is* the Type I error rate you are willing to tolerate.
- **Bigger samples** give narrower bootstrap intervals and more power — same data-generating reality, sharper picture.

Exercises

The most important Unit 2 skill is **choosing the right simulation before you run anything**. These exercises drill that decision. Answers are provided so you can check your work.

1. **Bootstrap or randomization test — and which one?** For each scenario, state **(i)** whether the goal is to *test* a claim or *estimate* a quantity, and **(ii)** the **specific procedure** (for example, "randomization test for a difference in proportions" or "bootstrap interval for one mean").

a. A nutritionist wants to estimate the proportion of cafeteria meals that meet a calorie guideline, with a margin of error.
b. Researchers ask whether a new store layout changes the *average* time shoppers spend in the store, comparing a treatment store to a control store.
c. A team estimates *how much higher* the average tip percentage is for servers who write a thank-you note versus those who don't.
d. A factory checks whether the defect rate on a line differs from the company target of 2%.
e. A pollster estimates the average number of hours students sleep, with uncertainty.
f. A trial asks whether the recovery rate (recovered: yes/no) is higher under a new drug than under a placebo.
g. Analysts estimate the difference in the proportion of repeat customers between two store designs.
h. A coach asks whether the team's mean sprint time is *faster than* the league benchmark of 12.0 seconds.

Show answer

	Goal	Procedure
a	estimate	**bootstrap** interval for **one proportion**
b	test	**randomization test** for a **difference in means**
c	estimate	**bootstrap** interval for a **difference in means**
d	test	**randomization test** for **one proportion** (vs. 0.02)
e	estimate	**bootstrap** interval for **one mean**
f	test	**randomization test** for a **difference in proportions**
g	estimate	**bootstrap** interval for a **difference in proportions**
h	test	**randomization test** for **one mean** (vs. 12.0)

The tell: *estimate / how much / with a margin* → **bootstrap**; *is there a difference / higher / changes* → **randomization test**. Then "one group vs. two groups" and "proportion vs. mean" pin down the exact version.

2. **What would you shuffle or resample?** For scenarios (b), (d), and (g) above, describe the *mechanical* step the simulation repeats thousands of times.

Show answer

(b) Randomization test — pool all shoppers' times, **shuffle the store labels** (treatment/control), and recompute the difference in mean times; repeat to build the null distribution. **(d)** Randomization test for one proportion — simulate samples of the same size *assuming* $p = 0.02$ (e.g., spin a 2%-defect spinner n times) and record each simulated defect proportion. **(g)** Bootstrap — **resample each store's customers with replacement** at the original group sizes and recompute the difference in repeat-customer proportions; repeat to build the sampling distribution of the difference.

3. **Spot the mismatch.** A student proposes a **bootstrap interval** to answer: *"Is there evidence that more than half of voters support the measure?"* Is that the right tool? If not, fix it.

   Show answer

   Mismatch. "Is there evidence that more than half…" is a **test** of a directional claim $(p > 0.5)$, not an estimate, so it calls for a **randomization test for one proportion** (against $p = 0.5$), reported as a p-value. (A one-sided bootstrap interval lying entirely above 0.5 *could* support the same conclusion, but the question as posed is a hypothesis test.)

4. **Which interpretation is correct?** A 95% bootstrap interval for a population mean is $(18.2, 21.6)$ minutes.

   i. There is a 95% probability the population mean is between 18.2 and 21.6.
   ii. About 95% of the data fall between 18.2 and 21.6.
   iii. If the whole study were repeated many times, about 95% of the intervals built this way would contain the true mean.

   Show answer

   (iii). Confidence is the long-run success rate of the *procedure*, not a probability about this one fixed interval (i), and not a statement about where individual values fall (ii).

5. **Interval ⇔ test.** A 90% randomization-based interval for a difference in means is $(-1.5, 4.0)$. Two-sided at $\alpha = 0.10$, would the matched test reject H_0 : difference $= 0$?

   Show answer

   No. The interval **contains 0**, so 0 stays plausible and the matched two-sided test does **not** reject at $\alpha = 0.10$. (A CI excluding the null value ⇔ the matched test rejects, at the corresponding level.)

6. **Errors in context.** A school will adopt a new reading program only if a test shows **discernible** evidence it beats the current one. Describe a **Type I** and a **Type II** error here, and one real consequence of each.

   Show answer

   Type I: conclude the new program is better when it truly is not → the school pays to switch for no gain. **Type II:** conclude there is no discernible benefit when the new program truly is better → students miss a real improvement. Choosing α controls the Type I rate; a larger sample lowers the Type II rate (raises power).

Revisit the engines. If any of this feels shaky, re-open the Randomization Test and Bootstrap tools and watch the null distribution and the sampling distribution form side by side — the contrast in the comparison table above is exactly what those two animations show.

Tech Tutorial: Simulate

Tech Tutorial — Simulate. This unit's tools build a sampling distribution by **resampling** — shuffling for a hypothesis test, bootstrapping for a confidence interval. Pick **one** tool and follow its tab.

- **StatLens** — purpose-built for resampling; the natural home for this unit.
- **R** — the `infer` package mirrors the same logic in code.
- **Jamovi** — its standard menus do *not* do randomization/bootstrap (see its tab).

We reuse the class survey from the Explore tutorial: `class_survey.csv`.

What you'll do

1. A **randomization test** — is mean `commute_min` different for On- vs. Off-campus students?
2. A **bootstrap confidence interval** — estimate the mean `sleep_hours`.

---

1. Randomization test for a difference in means

We ask whether on-campus and off-campus students differ in mean one-way commute. The **null** says housing doesn't matter, so any labeling of students as on/off-campus is interchangeable. We **shuffle** the housing labels many times to build the null distribution of the difference in means, then see where the observed difference falls.

10.8.1 StatLens

1. Open the Randomization: Difference in Means tool.
2. Set the **response** to `commute_min` and the **grouping** variable to `housing`.
3. Generate ~1000 shuffles. Read the **p-value** — the share of shuffles at least as extreme as the observed difference.

10.8.2 R

```
library(readr)
library(dplyr)
library(infer)
library(ggplot2)

survey <- read_csv("../datasets/class_survey.csv")

obs_diff <- survey |>
```

```
  specify(commute_min ~ housing) |>
  calculate(stat = "diff in means", order = c("Off-campus", "On-campus"))

null_dist <- survey |>
  specify(commute_min ~ housing) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "diff in means", order = c("Off-campus", "On-campus"))

null_dist |> get_p_value(obs_stat = obs_diff, direction = "two-sided")
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1       0
```

```
null_dist |>
  visualize() +
  shade_p_value(obs_stat = obs_diff, direction = "two-sided")
```

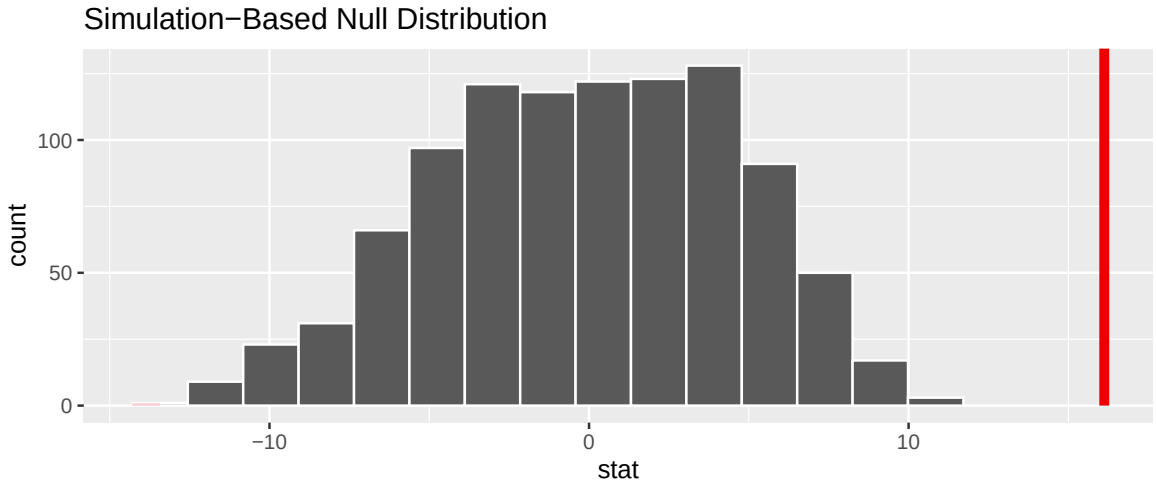


Figure 10.2: Randomization (null) distribution of the difference in mean commute. The observed difference sits far in the tail.

10.8.3 Jamovi

> **Jamovi: built-in menus don't shuffle**
>
> Jamovi's *built-in* menus don't run randomization (permutation) tests — a general permutation test on your data needs the **StatLens** tab (no install) or the **R** tab. (Add-on modules such as **esci** give menu-driven *sampling-variability* demos, but not a permutation p-value.) Jamovi returns in force in the next unit, *Calculate*, with the formula-based version of this comparison — a two-sample *t*-test.

2. Bootstrap confidence interval for a mean

To estimate the mean nightly sleep with a margin of error, we **bootstrap**: resample the data *with replacement* many times, record each resample's mean, and take the middle 95% of those means as the interval.

10.8.4 StatLens

1. Open the Bootstrap: One Mean tool.
2. Choose `sleep_hours`. Generate ~1000 bootstrap resamples.
3. Read the **95% confidence interval** (the 2.5th–97.5th percentiles).

10.8.5 R

```
boot_dist <- survey |>
  specify(response = sleep_hours) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "mean")

ci <- boot_dist |> get_confidence_interval(level = 0.95, type = "percentile")
ci
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1     6.82     7.32
```

```
boot_dist |>
  visualize() +
  shade_confidence_interval(endpoints = ci)
```

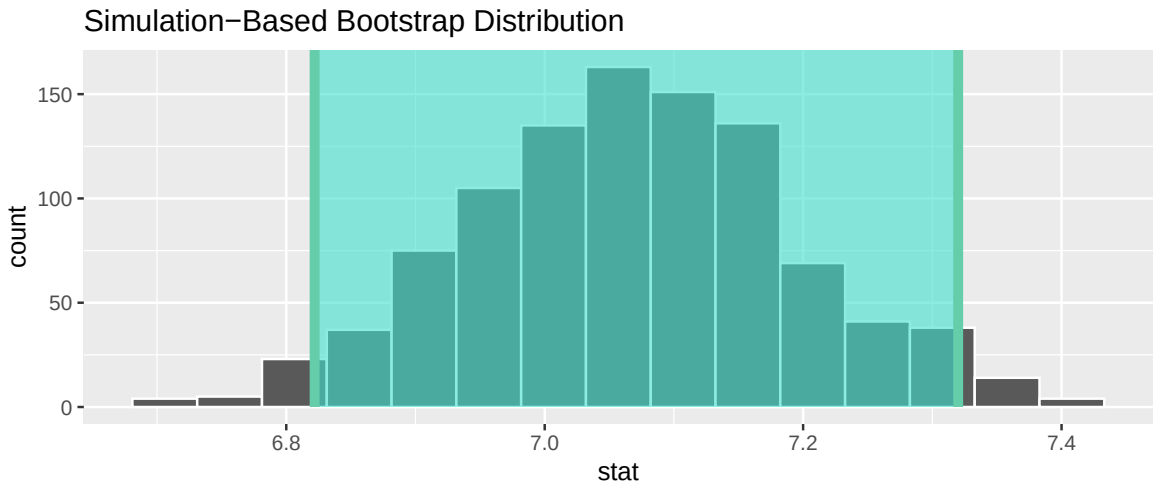


Figure 10.3: Bootstrap distribution of the mean sleep hours, with the 95% percentile interval shaded.

10.8.6 Jamovi

Jamovi: bootstrapping needs a module

A bootstrap CI for a plain mean isn't in Jamovi's built-in menus — use the **StatLens** or **R** tab. (The **esci** add-on does menu-driven interval estimation and the "dance of the means"; **GAMLj** does true bootstrap CIs for regression/ANOVA models.) In the *Calculate* unit, Jamovi produces the formula-based one-sample t interval for this same mean.

Check yourself

- ☐ run a randomization test for a difference and read its p-value;
- ☐ build and read a bootstrap confidence interval;
- ☐ explain what is being shuffled (test) vs. resampled with replacement (interval).

Going further in R (optional)

OpenIntro *IMS* interactive tutorials for this unit (same `infer` style as above):

- Sampling variability
- Randomization test
- Bootstrapping

PART III

UNIT 3: COMPUTE

You've built sampling distributions and computed confidence intervals through simulation. Now we introduce the mathematical tools — the normal distribution, the t-distribution — that approximate what simulation does. These theoretical methods are faster to compute and appear in published research, but they rest on the same logic you already understand.

Chapter 11

Normal Approximation

In the randomization tests and bootstrap confidence intervals chapters, we used simulation-based methods to make inferences about population parameters. Those methods are powerful and flexible, but they share a limitation: every new problem requires running thousands of simulations. In this chapter, we introduce the **normal distribution**, a mathematical model that describes the bell-shaped pattern we keep seeing in our sampling distributions. When certain conditions are met, the normal distribution lets us calculate probabilities and construct confidence intervals using formulas instead of simulations. This is not a replacement for simulation-based thinking – it is a complement. The normal model provides the theoretical justification for *why* those simulations work the way they do.

11.1 The normal distribution

Among all the distributions encountered in statistics, one is overwhelmingly the most common. The symmetric, unimodal, bell-shaped curve is so ubiquitous that it goes by many names: the **normal curve**, the **normal model**, or the **normal distribution**.

The **normal distribution** is a symmetric, unimodal, bell-shaped distribution described by two **parameters**: the mean μ (which determines the center) and the standard deviation σ (which determines the spread). We write the distribution in shorthand as $N(\mu, \sigma)$.

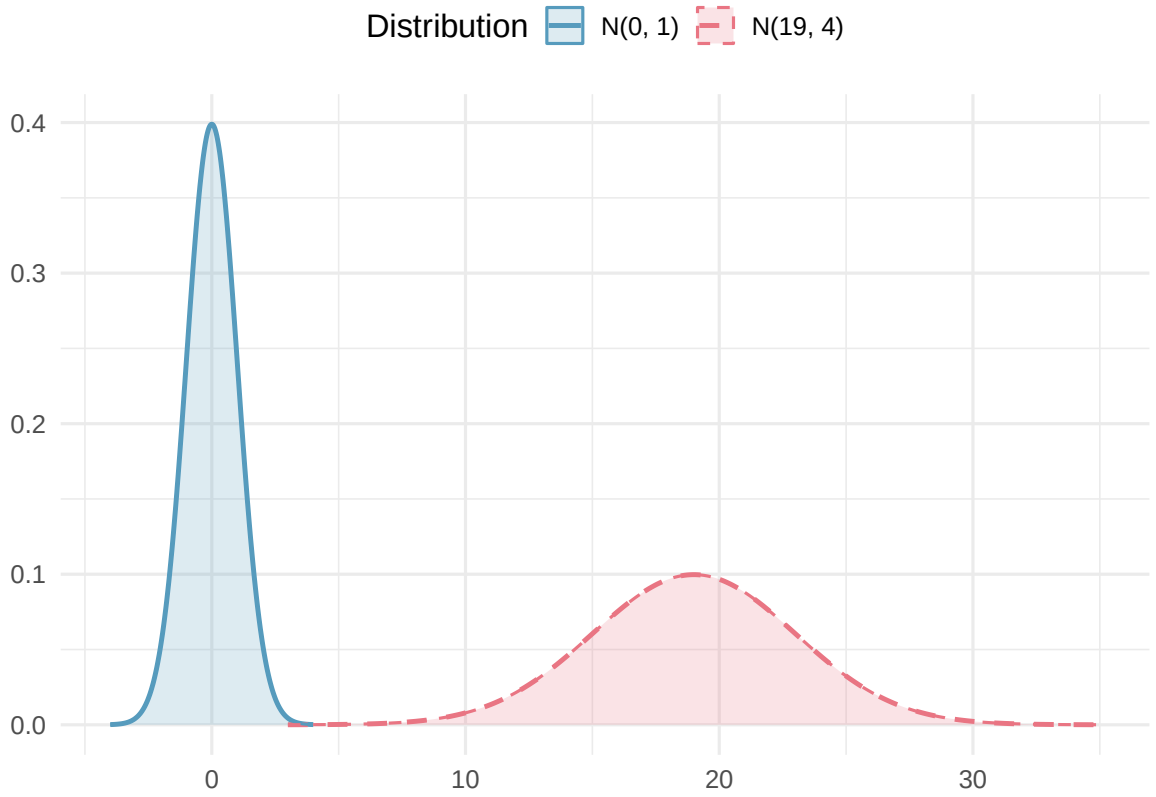


Figure 11.1: Two normal distributions with different parameters. The left curve has $\mu = 0$ and $\sigma = 1$; the right curve has $\mu = 19$ and $\sigma = 4$. Changing μ shifts the curve left or right; changing σ stretches or compresses it. ↗ Try this live — vary μ and σ and watch the curve slide and rescale.

Key properties of the normal distribution:

- The curve is perfectly symmetric about the mean μ.
- The total area under the curve is exactly 1 (representing 100% of observations). We can interpret any area under the curve as a **proportion** — and, equivalently, as the **probability** of a randomly selected observation falling in that range.
- The tails extend infinitely in both directions but get very close to zero far from the center.
- Changing μ shifts the entire curve left or right without changing its shape.
- Increasing σ makes the curve wider and shorter; decreasing σ makes it narrower and taller.

Normal distribution facts. Many variables are *nearly* normal, but none are *exactly* normal. The normal distribution, while not perfect for any single problem, is extremely useful for a wide variety of problems. Variables like SAT scores, adult heights, and blood pressure measurements closely follow the normal distribution.

11.1.1 The standard normal distribution

The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the **standard normal distribution**, written $N(0, 1)$. It serves as a reference distribution that we will use throughout the course.

Write down the shorthand for a normal distribution with each set of parameters.

a. Mean 5 and standard deviation 3

b. Mean -100 and standard deviation 10
c. Mean 2 and standard deviation 9

---

a. $N(\mu = 5, \sigma = 3)$
b. $N(\mu = -100, \sigma = 10)$
c. $N(\mu = 2, \sigma = 9)$

11.2 The 68-95-99.7 rule

One of the most useful facts about the normal distribution is how observations spread out around the mean. This relationship is captured by the **68-95-99.7 rule** (sometimes called the **empirical rule**).

The 68-95-99.7 rule. For any normal distribution $N(\mu, \sigma)$:

- About **68%** of observations fall within **1 standard deviation** of the mean: between $\mu - \sigma$ and $\mu + \sigma$.
- About **95%** of observations fall within **2 standard deviations** of the mean: between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- About **99.7%** of observations fall within **3 standard deviations** of the mean: between $\mu - 3\sigma$ and $\mu + 3\sigma$.

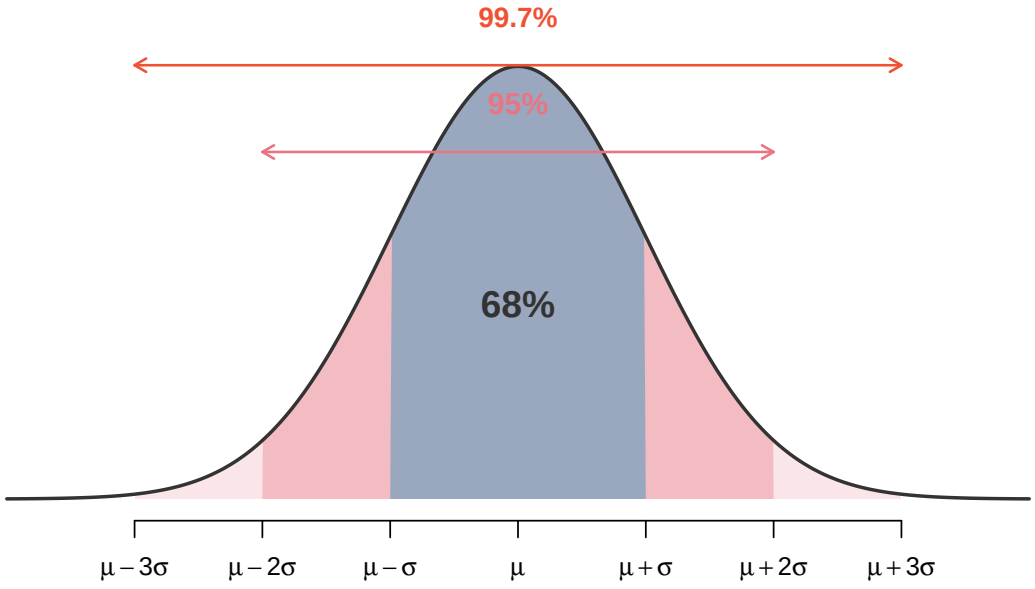


Figure 11.2: The 68-95-99.7 rule on a normal curve. About 68% of the area lies within $\mu \pm \sigma$, 95% within $\mu \pm 2\sigma$, and 99.7% within $\mu \pm 3\sigma$. Each labeled arrow bar spans exactly the region its percentage measures. ↗ Try this live — drag the cutlines to see the 68-95-99.7 boundaries live.

This rule gives quick, approximate answers without needing a calculator or software.

SAT scores closely follow the normal distribution with mean $\mu = 1500$ and standard deviation $\sigma = 300$.

a. About what percent of test takers score between 1200 and 1800?
b. About what percent score between 1500 and 2100?

---

a. 1200 is one standard deviation below the mean $(1500 - 300 = 1200)$, and 1800 is one standard deviation above the mean $(1500 + 300 = 1800)$. By the 68-95-99.7 rule, about **68%** of test takers score between 1200 and 1800.

b. The range from 900 to 2100 covers $\mu \pm 2\sigma$, which contains about 95% of scores. Since the normal distribution is symmetric, the range from 1500 to 2100 (the upper half) contains about $\frac{95\%}{2} = 47.5\%$ of scores.

Adult male heights in the US follow approximately a normal distribution with mean 70 inches and standard deviation 3.3 inches. Using the 68-95-99.7 rule, approximately what percentage of adult males are between 63.4 and 76.6 inches tall?

Show answer

63.4 inches is $70 - 2(3.3) = 63.4$ and 76.6 inches is $70 + 2(3.3) = 76.6$. These are each 2 standard deviations from the mean, so by the 68-95-99.7 rule, approximately 95% of adult males have heights in this range.

Continuing with SAT scores ($\mu = 1500$, $\sigma = 300$), about what percent of test takers score above 2100?

Show answer

2100 is 2 standard deviations above the mean. By the 68-95-99.7 rule, about 95% of scores are within 2 standard deviations, leaving about 5% outside. Since the distribution is symmetric, about 2.5% are above 2100.

11.3 Z-scores and standardizing

When comparing observations from different distributions, raw values are not directly comparable. A score of 1800 on the SAT and a score of 24 on the ACT look very different, but we need a common scale to determine which performance is better relative to other test takers. The solution is **standardizing** the observations.

The **Z-score** of an observation x from a distribution with mean μ and standard deviation σ is the number of standard deviations it falls above or below the mean:

$$Z = \frac{x - \mu}{\sigma}$$

- Positive Z-scores indicate observations **above** the mean.
- Negative Z-scores indicate observations **below** the mean.
- A Z-score of 0 means the observation equals the mean.

If the observation x comes from a normal distribution $N(\mu, \sigma)$, then its Z-score follows the standard normal distribution $N(0, 1)$. Standardizing shifts the center to 0 and rescales the spread to 1, but preserves the normal shape.

SAT scores follow $N(\mu = 1500, \sigma = 300)$ and ACT scores follow $N(\mu = 21, \sigma = 5)$. Nel scored 1800 on the SAT and Sian scored 24 on the ACT. Who performed better relative to other test takers?

We compute Z-scores for each student.

Nel's Z-score: $Z_{\text{Nel}} = \frac{1800-1500}{300} = \frac{300}{300} = 1.00$

Sian's Z-score: $Z_{\text{Sian}} = \frac{24-21}{5} = \frac{3}{5} = 0.60$

Nel is 1 standard deviation above the SAT mean, while Sian is only 0.6 standard deviations above the ACT mean. Nel performed better relative to other test takers.

Head lengths of brushtail possums follow a nearly normal distribution with mean 92.6 mm and standard deviation 3.6 mm. Compute the Z-scores for possums with head lengths of 95.4 mm and 85.8 mm. Which possum has the more unusual head length?

Show answer

For $x_1 = 95.4$ mm: $Z_1 = \frac{95.4-92.6}{3.6} = \frac{2.8}{3.6} = 0.78$. For $x_2 = 85.8$ mm: $Z_2 = \frac{85.8-92.6}{3.6} = \frac{-6.8}{3.6} = -1.89$. The second possum has a more unusual head length because $|-1.89| > |0.78|$.

Let X represent a value drawn at random from $N(\mu = 3, \sigma = 2)$, and suppose we observe $x = 5.19$.

a. Find the Z-score.
b. How many standard deviations above or below the mean is this observation?

---

a. $Z = \frac{x-\mu}{\sigma} = \frac{5.19-3}{2} = \frac{2.19}{2} = 1.095$

b. The observation is 1.095 standard deviations *above* the mean. We know it is above because the Z-score is positive.

11.4 Finding probabilities from the normal distribution

One of the most important skills in statistics is finding the probability that a normally distributed variable falls in a specified range. The idea is straightforward: **area under the curve equals probability**. You've already been doing this intuitively — in the randomization tests and bootstrap confidence intervals chapters, every time you counted "what fraction of simulated statistics were this extreme?" you were estimating a probability by looking at proportions in a distribution. The normal model just gives us a formula for the curve, so we can compute the area exactly instead of simulating it.

Strategy for normal probability problems. Always follow these steps:

1. **Draw a picture.** Sketch the normal curve, mark the mean, and shade the area you want to find.
2. **Find the Z-score.** Standardize the value(s) of interest using $Z = \frac{x-\mu}{\sigma}$.
3. **Look up the area.** Use a normal probability table, statistical software, or the StatLens Normal Distribution Explorer to find the area to the **left** of the Z-score.
4. **Adjust if needed.** If you need a right-tail area, subtract the left-tail area from 1. For an area between two values, subtract the smaller left-tail area from the larger.

11.4.1 Left-tail probabilities (percentiles)

The area to the left of a value on the normal curve gives the **percentile** – the proportion of observations that fall below that value.

The **percentile** of an observation is the percentage of values in the distribution that fall at or below that observation.

Nel scored 1800 on the SAT ($Z = 1.00$). What percentile is this?

We need the area to the left of $Z = 1.00$ on the standard normal curve.

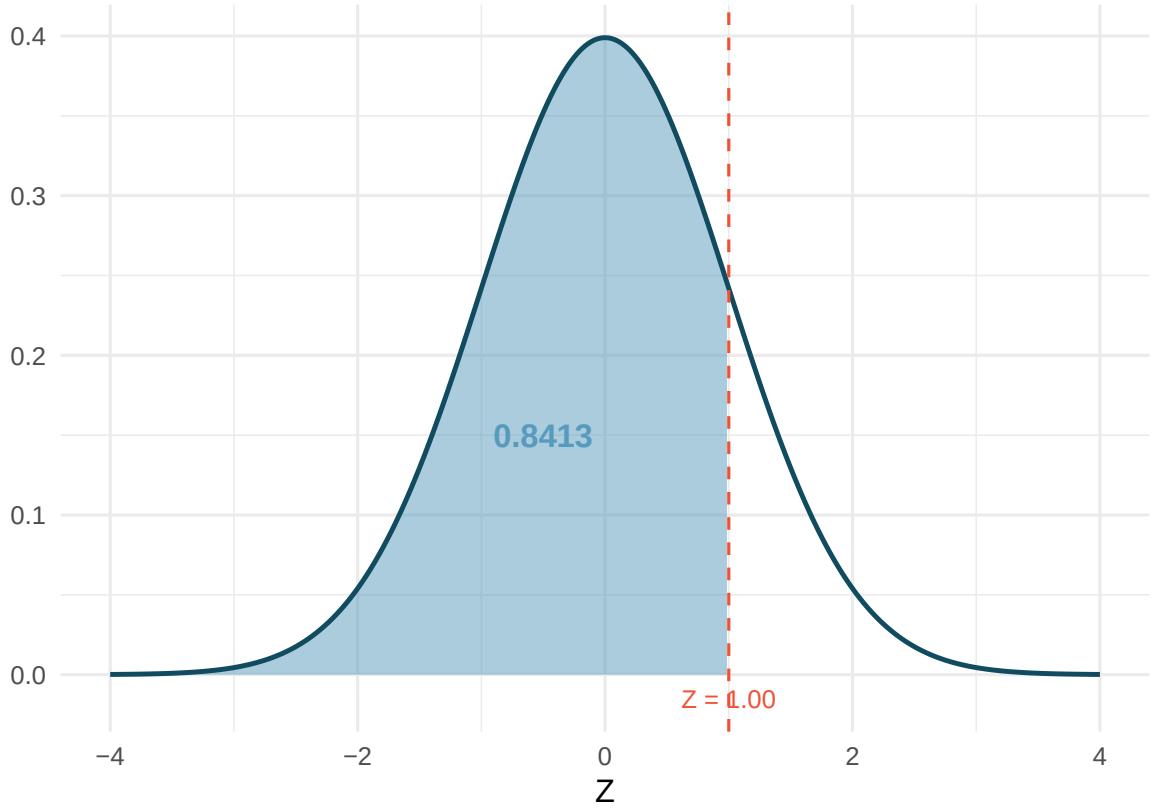


Figure 11.3: The standard normal curve with the area to the left of Z = 1.00 shaded. This area represents Nel's percentile. ↗ Try this live — drag the cutline to any Z value and read off the left-tail area.

Using a normal probability table or software, the area to the left of $Z = 1.00$ is **0.8413**. Nel is at the **84th percentile** – 84.13% of SAT takers scored lower.

11.4.2 Right-tail probabilities

Shannon is a randomly selected SAT taker. What is the probability Shannon scores at least 1630?

Step 1: Draw a picture. We shade the area above 1630 on the $N(1500, 300)$ curve.

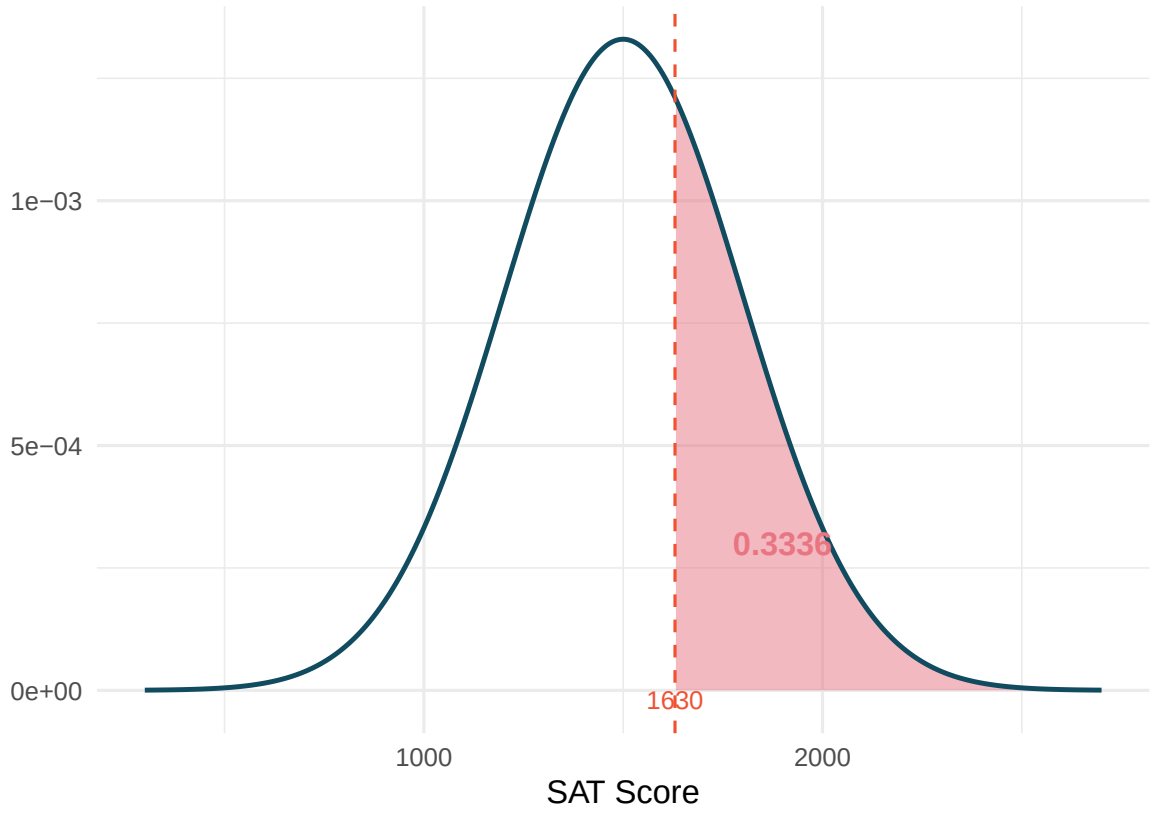


Figure 11.4: A normal curve with mean 1500 and standard deviation 300. The area to the right of 1630 is shaded. ↗ Try this live — set μ, σ, and the cutline to reproduce this area.

Step 2: Find the Z-score.

$$Z = \frac{1630 - 1500}{300} = \frac{130}{300} = 0.43$$

Step 3: Look up the left-tail area. The area to the left of $Z = 0.43$ is 0.6664.

Step 4: Adjust. The area to the *right* is $1 - 0.6664 = 0.3336$.

The probability Shannon scores at least 1630 is **0.3336**, or about 33.4%.

If the probability of Shannon scoring at least 1630 is 0.3336, what is the probability Shannon scores *less than* 1630?

Show answer

The probability of scoring less than 1630 is the area to the left: 0.6664, or about 66.6%. The left-tail and right-tail areas always sum to 1.

11.4.3 Probabilities between two values

Adult male heights follow $N(\mu = 70, \sigma = 3.3)$ inches. What is the probability that a randomly selected adult male is between 5′9″ (69 inches) and 6′2″ (74 inches)?

Step 1: Draw a picture. Shade the area between 69 and 74 on the normal curve.

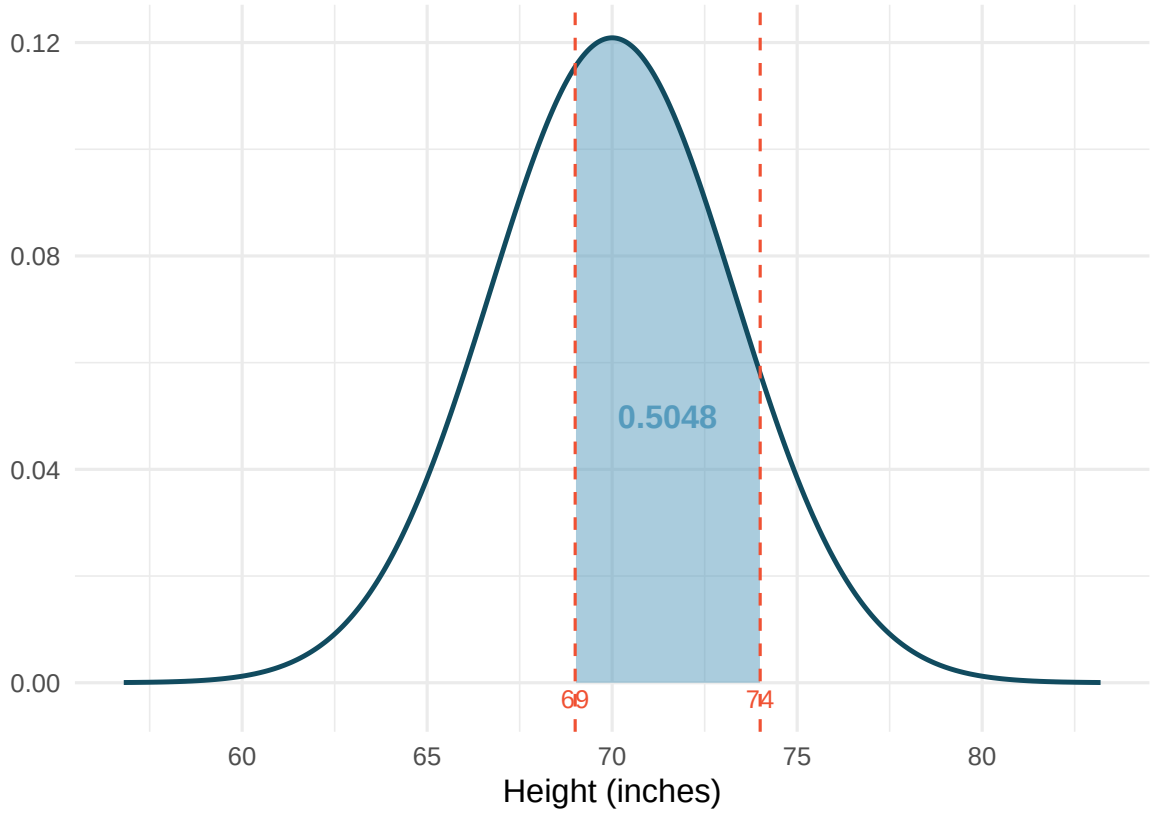


Figure 11.5: A normal curve centered at 70 with standard deviation 3.3. The area between 69 and 74 is shaded. ↗ Try this live — drag the two cutlines to shade any interval between them.

Step 2: Find both Z-scores.

$$Z_1 = \frac{69 - 70}{3.3} = -0.30 \qquad Z_2 = \frac{74 - 70}{3.3} = 1.21$$

Step 3: Look up both left-tail areas.

- Area to the left of $Z_1 = -0.30$: 0.3821
- Area to the left of $Z_2 = 1.21$: 0.8869

Step 4: Subtract.

$$P(69 < X < 74) = 0.8869 - 0.3821 = 0.5048$$

About **50.5%** of adult males are between 5′9″ and 6′2″.

What percent of SAT takers score between 1500 and 2000? (SAT scores follow $N(1500, 300)$.)

Show answer

First find the Z-scores: $Z_1 = \frac{1500-1500}{300} = 0.00$ and $Z_2 = \frac{2000-1500}{300} = 1.67$. The left-tail areas are 0.5000 and 0.9525. The area between them is $0.9525 - 0.5000 = 0.4525$, so about 45.25% score between 1500 and 2000.

11.4.4 Finding values from percentiles

Sometimes we want to work backwards: given a percentile, find the observation.

Adult male heights follow $N(70, 3.3)$. Yousef's height is at the 40th percentile. How tall is Yousef?

Step 1: Draw a picture. The 40th percentile means 40% of the area is to the left.

Step 2: Find the Z-score for the 40th percentile. Using a normal probability table (look for 0.40 in the body of the table) or software, $Z = -0.25$.

Step 3: Solve for x. Using the Z-score formula:

$$-0.25 = \frac{x - 70}{3.3}$$

$$x = 70 + (-0.25)(3.3) = 70 - 0.825 = 69.2 \text{ inches}$$

Yousef is approximately 69.2 inches (about 5'9") tall.

What is the 95th percentile for SAT scores?

Show answer

The Z-score for the 95th percentile is $Z = 1.645$. Then $x = 1500 + 1.645 \times 300 = 1500 + 493.5 = 1993.5$. The 95th percentile for SAT scores is approximately 1994.

11.4.5 The normal probability table

A **normal probability table** (also called a Z-table) lists areas to the left of Z-scores for the standard normal distribution. The table is organized with the ones and tenths digits of the Z-score in the rows, and the hundredths digit in the columns.

Common Z-scores and areas worth remembering:

Table 11.1: Commonly used Z-scores and their corresponding left-tail areas.

Z-score	Left-tail area	Interpretation
-1.96	0.0250	2.5th percentile
-1.645	0.0500	5th percentile
-1.00	0.1587	15.87th percentile
0.00	0.5000	50th percentile (the mean)
1.00	0.8413	84.13th percentile
1.645	0.9500	95th percentile
1.96	0.9750	97.5th percentile
2.576	0.9950	99.5th percentile

See it in action. Open the Normal Distribution Explorer to verify entries in the table above — enter a Z-score, check the shaded area, then work backwards from a percentile.

11.5 The normal approximation to sampling distributions

In the sampling distributions chapter, we used simulation to explore sampling distributions – the distributions of sample statistics computed from many repeated samples. A striking pattern emerged: regardless of whether we were looking at sample means, sample proportions, or differences in means or proportions, the sampling distributions tended to be **bell-shaped and symmetric**. This is not a coincidence.

11.5.1 The Central Limit Theorem

The Central Limit Theorem (CLT). When observations are independent and the sample size is sufficiently large, the sampling distribution of many common statistics (sample means, sample proportions, and their differences) is approximately normal, regardless of the shape of the population distribution.

Specifically:

- The sampling distribution of $\bar{x}$ is approximately $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.
- The sampling distribution of $\hat{p}$ is approximately $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

The CLT is one of the most remarkable results in all of mathematics. It explains why simulation-based null distributions and bootstrap distributions looked bell-shaped: they were approximating the normal distribution all along.

Two conditions for the CLT to apply:

1. **Independence.** The observations must be independent of one another. This is satisfied when data come from a random sample or a randomized experiment.
2. **Sample size.** The sample must be large enough. What counts as "large enough" depends on the context:
   - For **proportions**: at least 10 expected successes and 10 expected failures ($np \geq 10$ and $n(1-p) \geq 10$). This is called the **success-failure condition**.
   - For **means**: $n \geq 30$ is a common guideline, though smaller samples can work if the population is not strongly skewed.

11.5.2 Standard error: quantifying the variability of a statistic

The CLT tells us the sampling distribution is approximately normal. But how wide is that normal distribution? The answer is given by the **standard error**.

The **standard error (SE)** of a statistic is the standard deviation of its sampling distribution. It measures how much the statistic varies from sample to sample.

The standard error depends on the parameter of interest:

Table 11.2: Standard error formulas for common parameters.

Parameter	Point estimate	Standard error
Population mean μ	$\bar{x}$	$SE = \frac{s}{\sqrt{n}}$
Population proportion p	$\hat{p}$	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Difference in means $\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Parameter	Point estimate	Standard error
Difference in proportions $p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Margin of error. The distance $z^* \times SE$ is called the **margin of error**. For a 95% confidence interval, $z^* = 1.96$ (often rounded to 2), giving a margin of error of approximately $2 \times SE$. This is directly connected to the 68-95-99.7 rule: 95% of sample statistics fall within about 2 standard errors of the true parameter value.

11.5.3 Connecting simulation and normal approximation

Let's see the connection between simulation-based methods and the normal approximation in a concrete example. As a quick reminder from the randomization tests chapter, a hypothesis test weighs the **null hypothesis** H_0 ("no effect") against the **alternative** H_A ("real effect"), computes a p**-value** (probability of the observed data or more extreme under H_0), and **rejects** H_0 when the p-value falls below the discernibility level α.

In the randomization tests chapter, we tested whether opportunity cost reminders reduce spending. We simulated 1,000 randomization samples and found a null distribution that was bell-shaped with a center of 0. The observed difference was 0.20. To model that null distribution with a normal curve, we also need its spread: the standard deviation of the null distribution, which is the **standard error** (SE). You can estimate it by computing the standard deviation of the simulated differences; a formula we will meet in the inference for proportions chapter gives $SE \approx 0.078$, and we use that value here.

Simulation approach: Count how many of the 1,000 simulated differences were at least as extreme as 0.20. The proportion gives the p-value: 5 of the 1,000, or about 0.005.

Normal approximation approach: Model the null distribution as $N(0, 0.078)$. Compute the Z-score:

$$Z = \frac{0.20 - 0}{0.078} = 2.56$$

The area to the right of $Z = 2.56$ is 0.0052.

Both approaches give nearly the same p-value (0.005 vs. 0.0052) and the same conclusion: reject H_0. The normal approximation works because the conditions are met – independent observations and a sufficiently large sample.

11.5.4 Confidence intervals using the normal model

The normal approximation also lets us build confidence intervals without bootstrapping.

General confidence interval formula. When the sampling distribution of a statistic is approximately normal with standard error SE, a confidence interval takes the form:

$$\text{point estimate} \pm z^* \times SE$$

where z^* is chosen based on the desired confidence level:

Table 11.3: Critical values z^* for common confidence levels.

Confidence level	z^*
90%	1.645
95%	1.960
99%	2.576

A study examined whether stents reduce stroke risk. The observed difference in 30-day stroke rates (treatment minus control) was $\hat{p}_T - \hat{p}_C = 0.090$, with $SE = 0.028$. Construct a 95% confidence interval.

The conditions for the normal model have been verified. Using $z^* = 1.96$:

$$0.090 \pm 1.96 \times 0.028 = 0.090 \pm 0.055 = (0.035, 0.145)$$

We are 95% confident that implanting a stent increased the 30-day stroke rate by between 3.5 and 14.5 percentage points. Since the interval does not contain 0, we have evidence of a real difference.

11.6 When does the normal model apply – and when doesn't it?

The normal distribution is powerful but not universal. Here we summarize when it is and is not appropriate.

When the normal model is a good approximation:

- Sample proportions $\hat{p}$, when the success-failure condition is met ($np \geq 10$ and $n(1-p) \geq 10$).
- Sample means $\bar{x}$, when the sample size is at least 30 or the population is approximately normal (we will use the t-distribution for means – see the confidence intervals for means chapter).
- Differences in proportions or means, when both groups satisfy the above conditions.

When the normal model is NOT a good approximation:

- Small samples with a skewed population distribution.
- Proportions near 0 or 1 with small sample sizes (the success-failure condition fails).
- Data with strong dependence between observations.

A medical consultant claims that only 3 of their 62 liver transplant clients had complications, compared to the national rate of 10%. Is it appropriate to use the normal approximation to test whether this consultant's rate is lower?

Under the null hypothesis ($p_0 = 0.10$), we check the success-failure condition:

- Expected successes (complications): $np_0 = 62 \times 0.10 = 6.2$
- Expected failures (no complications): $n(1-p_0) = 62 \times 0.90 = 55.8$

Since the expected number of successes (6.2) is less than 10, the **success-failure condition is not met**. The normal approximation would underestimate the variability and give an inaccurate p-value. A simulation-based approach (as in the randomization tests chapter) would be more appropriate here.

When conditions fail, use simulation. If the success-failure condition is not met or the sample size is too small for the normal approximation, we should fall back on the simulation-based methods from

the randomization tests and bootstrap confidence intervals chapters. The bootstrap and randomization approaches work well even when the normal model does not. This is one of the great advantages of simulation-based inference: it is more broadly applicable.

11.7 Chapter review

11.7.1 Summary

This chapter introduced the normal distribution, the most commonly encountered distribution in statistics. Key concepts include:

- The normal distribution is described by two parameters: the mean μ (center) and standard deviation σ (spread), written as $N(\mu, \sigma)$.
- The **68-95-99.7 rule** provides quick estimates: about 68%, 95%, and 99.7% of observations fall within 1, 2, and 3 standard deviations of the mean.
- A **Z-score** standardizes an observation by measuring how many standard deviations it lies from the mean: $Z = \frac{x-\mu}{\sigma}$.
- Normal probabilities are found by computing Z-scores and looking up areas under the standard normal curve using tables, software, or StatLens.
- The **Central Limit Theorem** guarantees that sampling distributions of means and proportions are approximately normal when the sample is large enough and observations are independent.
- The **standard error** measures how much a statistic varies from sample to sample and is the key ingredient in confidence intervals and hypothesis tests.
- The normal approximation and simulation-based methods give similar results when conditions are met, but simulation is more flexible when conditions fail.

11.7.2 Key terms

Normal distribution, standard normal distribution, parameters (μ and σ), 68-95-99.7 rule, Z-score, standardizing, percentile, Central Limit Theorem (CLT), standard error (SE), margin of error, success-failure condition.

11.8 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Chronic illness.** In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions". However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used in this setting.

a. Create a 95% confidence interval for the proportion of U.S. adults who live with one or more chronic conditions. Also interpret the confidence interval in the context of the study. (Pew Research Center 2013)

b. Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

   i. We can say with certainty that the confidence interval from part (a) contains the true percentage of U.S. adults who suffer from a chronic illness.

ii. If we repeated this study 1,000 times and constructed a 95% confidence interval for each study, then approximately 950 of those confidence intervals would contain the true fraction of U.S. adults who suffer from chronic illnesses.

iii. The poll provides statistically discernible evidence (at the $\alpha = 0.05$ level) that the percentage of U.S. adults who suffer from chronic illnesses is below 50%.

iv. Since the standard error is 1.2%, only 1.2% of people in the study communicated uncertainty about their answer.

2. **Social media users and news, mathematical model.** A poll conducted in 2022 found that 50% of U.S. adults (i.e., a proportion of 0.5) get news from social media sometimes or often. The standard error for this estimate was 0.5% (i.e., 0.005), and a normal distribution may be used to model the sample proportion. (Pew Research Center 2022)

a. Construct a 99% confidence interval for the fraction of U.S. adults who get news on social media sometimes or often, and interpret the confidence interval in context.

b. Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

   i. The data provide statistically discernible evidence that more than half of U.S. adults users get news through social media sometimes or often. Use a discernibility level of $\alpha = 0.01$.

   ii. Since the standard error is 0.5%, we can conclude that 99.5% of all U.S. adults users were included in the study.

   iii. If we want to reduce the standard error of the estimate, we should collect less data.

   iv. If we construct a 90% confidence interval for the percentage of U.S. adults who get news through social media sometimes or often, the resulting confidence interval will be wider than a corresponding 99% confidence interval.

3. **Interpreting a Z score from a sample proportion.** Suppose that you conduct a hypothesis test about a population proportion and calculate the Z score to be 0.47. Which of the following is the best interpretation of this value? For the problems which are not a good interpretation, indicate the statistical idea being described.

a. The probability is 0.47 that the null hypothesis is true.

b. If the null hypothesis were true, the probability would be 0.47 of obtaining a sample proportion as far as observed from the hypothesized value of the population proportion.

c. The sample proportion is 0.47 standard errors greater than the hypothesized value of the population proportion.

d. The sample proportion is equal to 0.47 times the standard error.

e. The sample proportion is 0.47 away from the hypothesized value of the population.

f. The sample proportion is 0.47.

4. **Mental health.** The General Social Survey asked the question: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.

a. Interpret this interval in context of the data.

b. What does "95% confident" mean? Explain in the context of the application.

c. Suppose the researchers think a 99% confidence level would be more appropriate for this interval. Will this new interval be smaller or wider than the 95% confidence interval?

d. If a new survey were to be done with 500 Americans, do you think the standard error of the estimate be larger, smaller, or about the same.

5. **Area under the curve, Part I.** What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

a. $Z < -1.35$

b. $Z > 1.48$

c. $-0.4 < Z < 1.5$

d. $|Z| > 2$

6. **Area under the curve, Part II.** What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

a. $Z > -1.13$

b. $Z < 0.18$

c. $Z > 8$

d. $|Z| < 0.5$

7. **GRE scores, Part I.** Sophia who took the Graduate Record Examination (GRE) scored 160 on the Verbal Reasoning section and 157 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section for all test takers was 151 with a standard deviation of 7, and the mean score for the Quantitative Reasoning was 153 with a standard deviation of 7.67. Suppose that both distributions are nearly normal.

a. Write down the short-hand for these two normal distributions.

b. What is Sophia's Z-score on the Verbal Reasoning section? On the Quantitative Reasoning section? Draw a standard normal distribution curve and mark these two Z-scores.

c. What do these Z-scores tell you?

d. Relative to others, which section did she do better on?

e. Find her percentile scores for the two exams.

f. What percent of the test takers did better than her on the Verbal Reasoning section? On the Quantitative Reasoning section?

g. Explain why simply comparing raw scores from the two sections could lead to an incorrect conclusion as to which section a student did better on.

h. If the distributions of the scores on these exams are not nearly normal, would your answers to parts (b) - (f) change? Explain your reasoning.

8. **Triathlon times, Part I.** In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the *Men, Ages 30 - 34* group while Mary competed in the *Women, Ages 25 - 29* group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups. Can you help them? Here is some information on the performance of their groups:

- The finishing times of the *Men, Ages 30 - 34* group has a mean of 4313 seconds with a standard deviation of 583 seconds.
- The finishing times of the *Women, Ages 25 - 29* group has a mean of 5261 seconds with a standard deviation of 807 seconds.
- The distributions of finishing times for both groups are approximately Normal.

Remember: a better performance corresponds to a faster finish.

a. Write down the short-hand for these two normal distributions.

b. What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?

c. Did Leo or Mary rank better in their respective groups? Explain your reasoning.

d. What percent of the triathletes did Leo finish faster than in his group?

e. What percent of the triathletes did Mary finish faster than in her group?

f. If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning.

9. **GRE scores, Part II.** In a previous exercise we saw two distributions for GRE scores: $N(\mu = 151, \sigma = 7)$ for the verbal part of the exam and $N(\mu = 153, \sigma = 7.67)$ for the quantitative part. Use this information to compute each of the following:

a. The score of a student who scored in the 80th percentile on the Quantitative Reasoning section.

b. The score of a student who scored worse than 70% of the test takers in the Verbal Reasoning section.

10. **Triathlon times, Part II.** In a previous exercise we saw two distributions for triathlon times: $N(\mu = 4313, \sigma = 583)$ for *Men, Ages 30 - 34* and $N(\mu = 5261, \sigma = 807)$ for the *Women, Ages 25 - 29* group. Times are listed in seconds. Use this information to compute each of the following:

a. The cutoff time for the fastest 5% of athletes in the men's group, i.e. those who took the shortest 5% of time to finish.

b. The cutoff time for the slowest 10% of athletes in the women's group.

11. **LA weather, Part I.** The average daily high temperature in June in LA is 77° F with a standard deviation of 5° F. Suppose that the temperatures in June closely follow a normal distribution.

a. What is the probability of observing an 83° F temperature or higher in LA during a randomly chosen day in June?

b. How cool are the coldest 10% of the days (days with lowest high temperature) during June in LA?

12. **CAPM.** The Capital Asset Pricing Model (CAPM) is a financial model that assumes returns on a portfolio are normally distributed. Suppose a portfolio has an average annual return of 14.7% (i.e. an average gain of 14.7%) with a standard deviation of 33%. A return of 0% means the value of the portfolio doesn't change, a negative return means that the portfolio loses money, and a positive return means that the portfolio gains money.

a. What percent of years does this portfolio lose money, i.e. have a return less than 0%?

b. What is the cutoff for the highest 15% of annual returns with this portfolio?

13. **LA weather, Part II.** A previous exercise states that average daily high temperature in June in LA is 77° F with a standard deviation of 5° F, and it can be assumed that they follow a normal distribution. We use the following equation to convert ° F (Fahrenheit) to ° C (Celsius):

$$C = (F - 32) \times \frac{5}{9}.$$

a. Write the probability model for the distribution of temperature in ° C in June in LA.

b. What is the probability of observing a 28° C (which roughly corresponds to 83° F) temperature or higher in June in LA? Calculate using the ° C model from part (a).

c. Did you get the same answer or different answers in part (b) of this question and part (a) of a previous exercise? Are you surprised? Explain.

d. Estimate the IQR of the temperatures (in ° C) in June in LA.

14. **Find the SD (seedling heights).** Heights of seedlings in a greenhouse experiment are approximately normally distributed with mean 16 cm. The greenhouse classifies a seedling as "tall" when its height exceeds 20 cm, and about 15% of seedlings fall into that category.

a. What is the z-score corresponding to the "tall" cutoff?

b. Use the z-score and the cutoff to find the standard deviation of seedling heights.

c. Which step lets you solve for σ once the z-score is known?

15. **Overlay: normal curve on a simulated sampling distribution.** In a previous chapter you built the sampling distribution of $\widehat{p}$ by drawing many random samples from a population with proportion p and computing $\widehat{p}$ each time. Below are three simulated sampling distributions of $\widehat{p}$ for $p = 0.30$, using 5,000 samples each. Each histogram has a normal curve $N(\mu, \sigma)$ overlaid, with $\mu = p = 0.30$ and $\sigma = \sqrt{p(1-p)/n}$.

- **Panel A:** $n = 10$
- **Panel B:** $n = 50$
- **Panel C:** $n = 200$

(Imagine three histograms stacked, each with a smooth bell curve laid over it. At $n = 10$ the histogram is chunky and the curve doesn't hug it well, especially in the left tail near 0. At $n = 50$ the histogram is smoother and the curve fits well through the middle, with only small mismatch at the tails. At $n = 200$ the histogram and curve are nearly indistinguishable.)

a. In which panel does the normal curve fit the simulated sampling distribution best? Why?

b. In panel A ($n = 10$), name one visible reason the fit is poor. What condition from the chapter is failing?

c. The chapter states the success-failure condition ($np \geq 10$ and $n(1-p) \geq 10$) for the normal approximation. Check the condition for each panel using $p = 0.30$. Which panels satisfy it?

d. What does this tell you about *why* the normal approximation is a useful tool? What does it require to work?

16. **Bridge: simulation p-value vs. normal-approximation p-value.** In the randomization tests chapter, you tested for sex discrimination in a bank promotion study by shuffling promotion decisions many times and recording the difference in promotion rates between male and female applicants. The observed difference was $\hat{p}_M - \hat{p}_F = 0.292$. Suppose the randomization distribution of the difference, based on 10,000 shuffles, was centered near 0 with a standard deviation of about 0.13, and 245 of the 10,000 shuffles produced a difference ≥ 0.292.

    a. Compute the **simulation p-value** (one-tailed).

    b. Now compute the **normal-approximation p-value** for the same test. Treat the randomization distribution as approximately $N(0, 0.13)$ and compute the Z-score for the observed statistic 0.292; then use the tail probability under the standard normal (via technology).

    c. Do the two p-values agree? Within what tolerance?

    d. Under what conditions would you expect the two p-values to *disagree*? Name one situation where the simulation p-value would be more trustworthy.

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Read a normal probability in plain English.** Open the Normal Distribution tool preset to $\mu = 0$, $\sigma = 1$.
   a. Drag the boundary to $z = 1$. Report the area to the right. What is the matching area to the *left* of $z = -1$? Why are they equal?
   b. Switch to the "middle" tail and set the boundaries to ± 2. What is the area? Is your answer consistent with the **68–95–99.7 rule**?
   c. Switch back to "left tail" and set the boundary so the area is exactly 0.025. What z-value did you get? Where will you use this number later in the course?
   d. A student says: "Areas under the curve are probabilities, so they must be less than 1." Could you have an area of 1? An area of 0? What do those represent?

2. **Standardize, then read off the probability.** SAT math scores are approximately $N(\mu = 520, \sigma = 110)$.
   a. Compute the z-score for a student who scored 700.
   b. Use Normal at $\mu = 520, \sigma = 110$ to find the proportion of test-takers scoring above 700.
   c. Now use the *standardized* normal at $\mu = 0, \sigma = 1$ and set the boundary to your z-score from (a). Same answer?
   d. A counselor says: "About 5% of students score 700 or higher." Explain in one sentence what is *probabilistic* about the 5% and what is *not*.

3. **The sampling distribution of $\bar{x}$ is also normal — with different σ.** Continuing the SAT example: a school selects a random sample of $n = 25$ students and computes their mean math score $\bar{x}$.
   a. What is the sampling distribution of $\bar{x}$? Give its mean, standard error, and shape.
   b. Use the appropriate normal tool (pre-loaded with $\mu = 520$, $\sigma = 110/\sqrt{25} = 22$) to find the probability that $\bar{x} \geq 550$.
   c. Compare this to the probability that an **individual student** scores ≥ 550. Which is bigger and why?
   d. The school doubles n to 50. Will the probability in (b) get bigger or smaller? Why?

4. **Normal approximation to a proportion — when does the formula work?** Open the One-Proportion z-Test tool.
   a. Enter a scenario where $\hat{p}$ is close to 0.5 and n is large (say $\hat{p} = 0.45, n = 200$). Run a test. The tool reports the success–failure counts. Does the normal approximation apply?
   b. Now enter a scenario with small $\hat{p}$: $\hat{p} = 0.05$, $n = 40$. Does the normal approximation apply? Why or why not?
   c. For the small-$\hat{p}$ scenario in (b), if the test gives a small p-value, should you trust it? What should you do instead?
   d. Why is the success–failure rule based on np_0 and $n(1-p_0)$ (using the **null** p_0) rather than $n\hat{p}$ and $n(1-\hat{p})$ for a hypothesis test?

5. **Simulation vs. analytic — when does the normal model agree with the bootstrap?** Pair the One-Proportion z-Test with the Bootstrap CI for One Proportion on two datasets.

   First a well-behaved case: any dataset with $\hat{p}$ near 0.5 and $n \geq 100$.

   a. Compare the 95% intervals from the two tools. How close are they?

   Now a troublesome case: a dataset with small $\hat{p}$ (e.g., the medical consultant data, $\hat{p} \approx 0.05$).

b. Compare the 95% intervals again. Where do they disagree — in the midpoint, the width, or both?
c. Look at the **bootstrap distribution** of $\hat{p}$. Is it symmetric or skewed? How does that explain the disagreement with the symmetric z-interval?
d. A student says: "Both intervals are 'right' — they're just approximations." What's a more accurate framing? Which interval is more honest in the troublesome case, and what does "honest" mean for an interval?

Chapter 12

Inference for Proportions

In earlier chapters, we used simulation-based methods to draw conclusions about population proportions. In the randomization tests and bootstrap confidence intervals chapters, we built randomization distributions and bootstrap confidence intervals; in the normal approximation chapter, we learned that the normal distribution provides a mathematical model for these sampling distributions when certain conditions are met. In this chapter, we bring everything together for proportions. We develop normal-based confidence intervals and hypothesis tests for a single proportion and for the difference between two proportions. Throughout, we compare these formula-based results to the simulation-based results from earlier chapters, reinforcing the idea that these are two routes to the same destination.

12.1 Case study: is the complication rate lower?

People providing an organ for donation sometimes seek the help of a special **medical consultant**. These consultants assist the patient in all aspects of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery.

One consultant tried to attract patients by noting the national complication rate for liver donor surgeries is about 10%, but her clients have had only 3 complications in the 62 liver donor surgeries she has facilitated. She claims this is strong evidence that her work reduces complications.

Let p denote the true complication rate for this consultant's clients. The point estimate is

$$\hat{p} = \frac{3}{62} = 0.048.$$

The sample rate 4.8% is well below the national 10%. But we saw the same puzzle in the sampling variability chapter: even when nothing is going on, sample proportions vary from sample to sample. **Could a rate as low as 4.8% out of 62 have happened by chance, if the true rate really were 10%?** That is the question a hypothesis test is built to answer:

- H_0: The consultant's true complication rate matches the national rate, $p = 0.10$.
- H_A: The consultant's true complication rate is lower, $p < 0.10$.

To weigh the evidence, we need to know how much $\hat{p}$ would vary from sample to sample **if H_0 were true**. That variability is captured by the null distribution of $\hat{p}$.

12.2 Building the null distribution by simulation

We do not need calculus or a formula to picture the null distribution — we can build it directly with simulation, just as we did in the randomization tests chapter. But this is a randomization test of a *different kind*. There we compared two groups by **shuffling the group labels** on the data we already had; here there is only **one** group and a single claimed value ($H_0 : p = 0.10$), so there are no labels to shuffle. Instead we **draw fresh samples from a null population** built to match H_0 — a world in which the true complication rate really is 10%. Sampling repeatedly from a hypothesized population this way is exactly the **sampling-distribution** idea that the normal approximation will formalize in the next section; the simulation lets us *see* it first, concretely, before any formula.

12.2.1 The marbles metaphor

Imagine a large bag of marbles: 10% orange (complication) and 90% blue (no complication). This bag models the world under H_0. Each surgery is one draw from the bag; a "study" of $n = 62$ surgeries is one scoop of 62 marbles. The proportion of orange marbles in the scoop is one simulated $\hat{p}_{sim}$ from the null distribution.

If we scoop 62 marbles and get 5 orange and 57 blue, then

$$\hat{p}_{sim1} = \frac{5}{62} = 0.081.$$

That is *one* simulated study. Under H_0, this could easily have happened, even though it does not equal 0.10 exactly.

Is one simulated $\hat{p}_{sim}$ enough to draw a conclusion about the consultant?

No. One draw tells us nothing about the *shape* of the null distribution — it is a single value that happens to land somewhere by chance. To know what values of $\hat{p}$ are typical and which are extreme under H_0, we need many simulations.

12.2.2 From one draw to many

What happens as we accumulate more draws — does the null distribution take a recognizable shape? Rather than jump straight from one draw to a smooth histogram, it helps to build the distribution a few increments at a time.

Predict → Do → Explain. Open the Sampling Distribution Lab (proportion mode) — preloaded with $p = 0.10$ and $n = 62$, the null-hypothesis world for the medical-consultant case. The gated activity opens with a **Predict** step (commit to where the distribution of $\hat{p}_{sim}$ will be centered, how spread out, and what shape), walks you through drawing 1, 10, 100, and 1,000 samples in sequence, and closes with an **Explain** step where you articulate what the buildup showed. Your responses save locally in the tool, so you can return to your work on the same device.

Three ideas the activity lands on — worth flagging in the reading so you can compare them to your own conclusions:

- More samples do **not** make the individual $\hat{p}_{sim}$ values more precise — they fill in the *same* null distribution. To get narrower $\hat{p}_{sim}$ values you must increase n (the size of each scoop), not the number of scoops.
- The null distribution has a stable center at $p = 0.10$ and a stable spread even while the individual draws stay noisy. That stability is what makes the p-value calculation possible.

- The distribution is **roughly bell-shaped, but with a mild right skew**. That skew is not noise — it is a signal. With $p = 0.10$ and $n = 62$, the expected number of successes is $np = 6.2 < 10$, so the success–failure condition fails and the normal approximation is only approximate. When p sits near 0 (or 1) and n is not large, $\hat{p}$ has more room to wander toward the middle than toward the floor, which pushes a tail out to the right. This is the case where a simulated null distribution is more trustworthy than a normal formula — exactly the situation we started the chapter with.

12.2.3 The null distribution at 1,000 simulations

A computer can repeat the marbles scoop as many times as we want. Here are the results of 1,000 simulated studies under H_0: each study draws 62 marbles from the 10% / 90% bag, and we record $\hat{p}_{sim}$ from each.

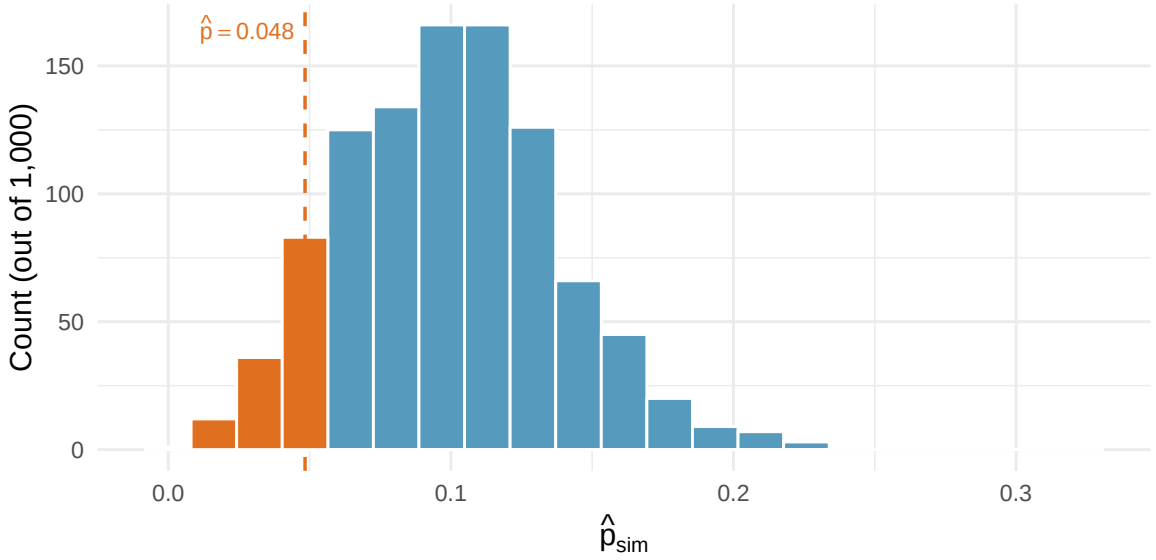


Figure 12.1: The null distribution of p-hat from 1,000 simulated studies under H□: p = 0.10, n = 62. Simulations with p-hat ≤ 0.048 (as extreme as the consultant's observed value, or more extreme) are shaded orange. ↗ Try this live — simulate under H_0 with your own p and n and rebuild the null.

As the bullets above anticipated, the distribution is **centered at** $p = 0.10$ (because we built it under H_0) and is roughly bell-shaped with a mild right skew (because $p = 0.10$ sits near 0 and $\hat{p}$ cannot go below it). Now we can read the evidence.

12.2.4 Reading the p-value

The consultant's observed $\hat{p} = 0.048$ is marked with the dashed line. The orange-shaded bars are the simulated $\hat{p}_{sim}$ values that are *at least as extreme* as the observed value in the direction of the alternative — here, $\hat{p}_{sim} \leq 0.048$.

The p-value is the fraction of the simulated studies that landed in the shaded tail:

$$p\text{-value} = \frac{\text{simulations with } \hat{p}_{sim} \leq 0.048}{1{,}000} \approx 0.132.$$

Using $\alpha = 0.05$, what conclusion should we draw?

The p-value ≈ 0.132 is greater than $\alpha = 0.05$. We **fail to reject** H_0. The data do not provide convincing evidence that the consultant's true complication rate is lower than the national 10% — a rate of 3 in 62 could plausibly occur even if her true rate were 10%.

Failing to reject H_0 is not the same as concluding that her rate *equals* 10%. It only says the data are compatible with the national rate; a larger study would be needed to distinguish a small real improvement from chance.

Simulating the null distribution of $\hat{p}$.

Whenever we want to test $H_0 : p = p_0$ but a normal-based formula is unreliable, we can build the null distribution directly:

1. Set up a marbles-bag with the proportion p_0 specified by H_0.
2. Repeatedly scoop n marbles and record the simulated $\hat{p}_{sim}$.
3. Read the p-value as the tail area of the null distribution beyond the observed $\hat{p}$.

This works for **any** sample size and does not require the success–failure condition. Because this is a *single* proportion, the exact null probabilities are in fact given by the **binomial distribution** (see the Binomial Distribution chapter) — the simulation is a tangible stand-in for that exact calculation, and the normal formula in the next section is a fast approximation to it. We lead with the simulation because *sampling from the null world* is the idea at the heart of every hypothesis test; the exact and approximate formulas come after.

12.3 Normal approximation for proportions

12.3.1 The sampling distribution of $\hat{p}$

In the normal approximation chapter, we saw that the Central Limit Theorem guarantees that the sampling distribution of $\hat{p}$ is approximately normal when certain conditions are met. Let's state these conditions precisely.

Sampling distribution of $\hat{p}$.

The sampling distribution of the sample proportion $\hat{p}$, based on a sample of size n from a population with true proportion p, is approximately normal when:

1. **Independence.** The observations are independent (e.g., from a simple random sample).
2. **Success-failure condition.** We expect to see at least 10 successes and 10 failures: $np \geq 10$ and $n(1-p) \geq 10$.

When these conditions are met, the sampling distribution is approximately:

$$\hat{p} \sim N\left(p,\ \sqrt{\frac{p(1-p)}{n}}\right)$$

The **success-failure condition** requires that the expected number of successes (np) and the expected number of failures ($n(1-p)$) are both at least 10. This ensures the sampling distribution of $\hat{p}$ is not too skewed for the normal approximation to work well.

A medical consultant facilitated 62 liver transplant surgeries, with 3 complications. Under the null hypothesis that the true complication rate is $p_0 = 0.10$, is it appropriate to use the normal approximation?

Check the success-failure condition using $p_0 = 0.10$:

- Expected successes (complications): $np_0 = 62 \times 0.10 = 6.2$
- Expected failures: $n(1 - p_0) = 62 \times 0.90 = 55.8$

The expected number of successes (6.2) is less than 10. The **success-failure condition is not met**, so the normal approximation should not be used here. A simulation-based approach (such as the null-distribution simulation earlier in this chapter) would be more appropriate.

A poll surveys 826 randomly selected payday loan borrowers and finds that 51% support a new regulation. Check whether the normal approximation can be used to model $\hat{p}$ for a hypothesis test with $H_0 : p = 0.50$.

Show answer

Check the success-failure condition using $p_0 = 0.50$: $np_0 = 826 \times 0.50 = 413$ and $n(1 - p_0) = 826 \times 0.50 = 413$. Both are well above 10, so the success-failure condition is met. Combined with the random sample (independence), the normal approximation is appropriate.

12.3.2 The standard error of $\hat{p}$

Standard error of $\hat{p}$.

When conditions are met, the variability of $\hat{p}$ is described by:

$$SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Since we rarely know the true p:

- **For confidence intervals**, use $\hat{p}$ as the best guess: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- **For hypothesis tests**, use p_0 (the value from H_0): $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

Consider polls of size 300 where approximately 2/3 of voters support legalized marijuana. Calculate the standard error of $\hat{p}$.

Show answer

Using $p \approx 2/3$: $SE = \sqrt{\frac{(2/3)(1/3)}{300}} = \sqrt{\frac{2/9}{300}} = \sqrt{0.000741} = 0.027$.

12.4 Confidence interval for a single proportion

12.4.1 The formula

Confidence interval for a single proportion p.

$$\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z^* is the critical value for the desired confidence level (e.g., $z^* = 1.96$ for 95%).

Steps:

1. Check conditions: independence and success-failure using $\hat{p}$.
2. Compute SE using $\hat{p}$.
3. Apply the confidence interval formula.

12.4.2 Worked example: Payday loan regulation

A simple random sample of 826 payday loan borrowers was surveyed, and 70% supported new regulations on payday lenders. Construct a 95% confidence interval for the true proportion who support the regulations.

Step 1: Check conditions.

- *Independence.* The data are a random sample.
- *Success-failure condition.* Using $\hat{p} = 0.70$: $n\hat{p} = 826 \times 0.70 = 578$ and $n(1-\hat{p}) = 826 \times 0.30 = 248$. Both are well above 10.

Step 2: Compute SE.

$$SE = \sqrt{\frac{0.70 \times 0.30}{826}} = \sqrt{\frac{0.21}{826}} = \sqrt{0.000254} = 0.016$$

Step 3: Construct the interval.

$$0.70 \pm 1.96 \times 0.016 = 0.70 \pm 0.031 = (0.669, 0.731)$$

We are 95% confident that between 66.9% and 73.1% of all payday loan borrowers support the new regulation.

12.4.3 Changing the confidence level

The confidence level is controlled by the choice of z^*:

Table 12.1: Critical values for common confidence levels.

Confidence level	z^*	Interval width
90%	1.645	Narrower
95%	1.960	Moderate
99%	2.576	Wider

Higher confidence requires a wider interval – we must cast a wider net to be more confident we have captured the true parameter.

Using the payday loan data ($\hat{p} = 0.70$, $n = 826$, $SE = 0.016$), construct a 99% confidence interval.

Show answer

Using $z^* = 2.576$: $0.70 \pm 2.576 \times 0.016 = 0.70 \pm 0.041 = (0.659, 0.741)$. The 99% interval is wider than the 95% interval, as expected.

A study on stent surgery found a point estimate of 0.090 (the difference in stroke rates) with $SE = 0.028$. Construct a 90% confidence interval.

Show answer

Using $z^* = 1.645$: $0.090 \pm 1.645 \times 0.028 = 0.090 \pm 0.046 = (0.044, 0.136)$. We are 90% confident that stent implantation increases the 30-day stroke rate by between 4.4% and 13.6%.

General confidence interval formula. If a point estimate follows a normal model with standard error SE, then a confidence interval for the population parameter is:

$$\text{point estimate} \pm z^* \times SE$$

where z^* corresponds to the chosen confidence level.

Planning a study: choosing n

Before collecting data, researchers often need to decide how large a sample to draw. The width of the CI above is driven by the standard error $\sqrt{\hat{p}(1-\hat{p})/n}$, so hitting a target margin of error is really a question of choosing n large enough. We work through the arithmetic (with worst-case $\hat{p} = 0.5$ and other refinements) in the sample size chapter later in the book.

12.5 Hypothesis test for a single proportion

12.5.1 The Z test for a proportion

The test statistic for a single proportion is a Z.

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

where $\hat{p}$ is the sample proportion and p_0 is the hypothesized proportion from H_0.

When H_0 is true and conditions are met, Z follows a standard normal distribution $N(0, 1)$.

Conditions:

- Independent observations
- $np_0 \geq 10$ and $n(1-p_0) \geq 10$

Note that for hypothesis tests, the standard error is computed using p_0 (not $\hat{p}$), because we are evaluating the data under the assumption that H_0 is true.

12.5.2 Worked example: Support for credit checks

From a random sample of 826 payday loan borrowers, 51% said they would support a regulation requiring lenders to pull credit reports. Is there convincing evidence that a majority (more than 50%) of borrowers support this regulation?

Step 1: State hypotheses.

- H_0: $p = 0.50$ (no majority support)
- H_A: $p > 0.50$ (majority support)

This is a one-sided test.

Step 2: Check conditions.

- *Independence.* Random sample.
- *Success-failure.* Using $p_0 = 0.50$: $np_0 = 826 \times 0.50 = 413$ and $n(1-p_0) = 413$. Both ≥ 10.

Step 3: Compute the test statistic and p-value.

$$SE = \sqrt{\frac{0.50 \times 0.50}{826}} = \sqrt{\frac{0.25}{826}} = 0.017$$

$$Z = \frac{0.51 - 0.50}{0.017} = \frac{0.01}{0.017} = 0.59$$

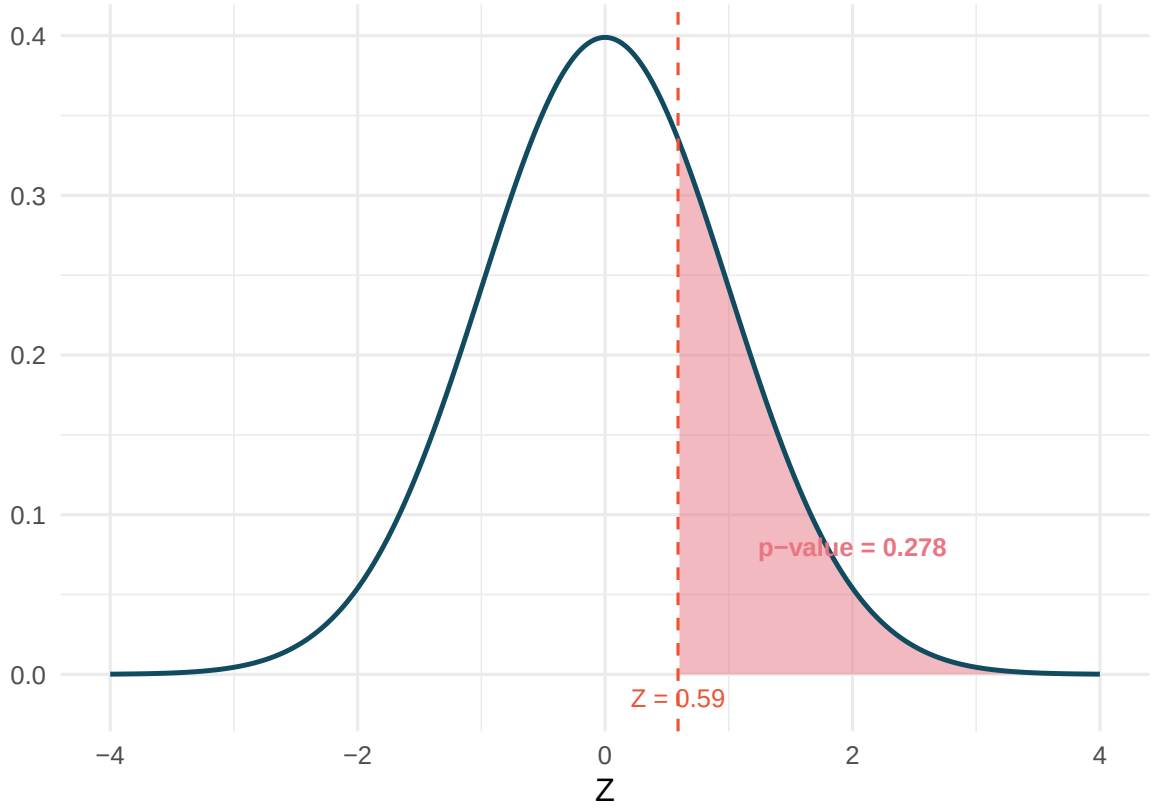


Figure 12.2: A standard normal curve with the area to the right of Z = 0.59 shaded, representing the one-sided p-value. ↗ Try this live — drag the cutline to any Z value and read off the right-tail area.

The area to the right of $Z = 0.59$ is 0.278.

Step 4: Conclusion. The p-value of 0.278 is much larger than $\alpha = 0.05$, so we fail to reject H_0. The poll does not provide convincing evidence that a majority of payday loan borrowers support the regulation requiring credit checks.

A candidate claims that more than 60% of voters in a district support her. A random sample of 500 voters finds $\hat{p} = 0.64$. Test the claim at $\alpha = 0.05$.

Show answer

Hypotheses: $H_0 : p = 0.60$, $H_A : p > 0.60$. Success-failure: $500(0.60) = 300$ and $500(0.40) = 200$, both ≥ 10. $SE = \sqrt{0.60 \times 0.40/500} = \sqrt{0.000480} = 0.0219$. $Z = (0.64 - 0.60)/0.0219 = 1.83$. The right-tail area is 0.034. Since $0.034 < 0.05$, we reject H_0. There is sufficient evidence that more than 60% of voters support the candidate.

12.5.3 When conditions fail

When the success-failure condition is not met, the normal model can underestimate the variability of $\hat{p}$, leading to inaccurate p-values and confidence intervals. In these cases, use simulation-based methods:

- **Simulation-based hypothesis test**: Simulate samples under H_0 by drawing from a population with proportion p_0 (the null-distribution simulation from the start of this chapter).
- **Bootstrap confidence interval**: Resample from the original data to estimate the variability of $\hat{p}$.

12.6 Difference of two proportions

12.6.1 The parameter and point estimate

When comparing two groups, the parameter of interest is the difference in population proportions: $p_1 - p_2$. The point estimate is $\hat{p}_1 - \hat{p}_2$.

12.6.2 Confidence interval for a difference in proportions

Confidence interval for $p_1 - p_2$.

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Conditions:

1. **Independence (extended).** Observations are independent within and between groups.
2. **Success-failure condition (each group separately).** $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1-\hat{p}_2) \geq 10$.

In a CPR study, patients who had a heart attack were randomly assigned to receive a blood thinner (treatment) or not (control). The survival rates after 24 hours:

Table 12.2: CPR study results.

Group	n	Survived	$\hat{p}$
Treatment	40	14	0.350
Control	50	11	0.220

Construct a 90% confidence interval for the difference in survival rates, $p_T - p_C$.

Check conditions.

- *Independence.* Patients were randomly assigned, so groups are independent.
- *Success-failure.* Treatment: $40(0.35) = 14 \geq 10$ and $40(0.65) = 26 \geq 10$. Control: $50(0.22) = 11 \geq 10$ and $50(0.78) = 39 \geq 10$. All conditions met.

Compute the interval.

$$\hat{p}_T - \hat{p}_C = 0.350 - 0.220 = 0.130$$

$$SE = \sqrt{\frac{0.350(0.650)}{40} + \frac{0.220(0.780)}{50}} = \sqrt{\frac{0.2275}{40} + \frac{0.1716}{50}} = \sqrt{0.005688 + 0.003432} = \sqrt{0.009120} = 0.095$$

Using $z^* = 1.645$ for 90% confidence:

$$0.130 \pm 1.645 \times 0.095 = 0.130 \pm 0.156 = (-0.026, 0.286)$$

We are 90% confident that the blood thinner changes the 24-hour survival rate by somewhere between -2.6 and $+28.6$ percentage points. Because the interval contains 0, we cannot conclude that the blood thinner affects survival rates. But the interval also contains sizable benefits — up to about 29 percentage points — so this small study cannot rule out a meaningful effect either.

12.6.3 Hypothesis test for a difference in proportions

The test statistic for comparing two proportions is a Z.

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where the **pooled proportion** is:

$$\hat{p}_{pool} = \frac{\text{total successes in both groups}}{\text{total sample size}} = \frac{x_1 + x_2}{n_1 + n_2}$$

When H_0 is true ($p_1 = p_2$), both groups share the same population proportion, so we pool the data to get a single best estimate.

Conditions:

- Independence within and between groups.
- $n_1\hat{p}_{pool} \geq 10$, $n_1(1 - \hat{p}_{pool}) \geq 10$, $n_2\hat{p}_{pool} \geq 10$, $n_2(1 - \hat{p}_{pool}) \geq 10$.

Does the blood thinner improve 24-hour survival after CPR? Using the data from the previous example (treatment: 14/40 survived; control: 11/50 survived), test at $\alpha = 0.05$.

Step 1: State hypotheses.

- H_0: $p_T - p_C = 0$ (the blood thinner has no effect)
- H_A: $p_T - p_C \neq 0$ (the blood thinner has some effect)

Step 2: Check conditions.

The pooled proportion is $\hat{p}_{pool} = \frac{14+11}{40+50} = \frac{25}{90} = 0.278$. Check:

- Treatment: $40(0.278) = 11.1 \geq 10$, $40(0.722) = 28.9 \geq 10$
- Control: $50(0.278) = 13.9 \geq 10$, $50(0.722) = 36.1 \geq 10$

Conditions are met.

Step 3: Compute the test statistic and p-value.

$$SE = \sqrt{0.278(0.722)\left(\frac{1}{40} + \frac{1}{50}\right)} = \sqrt{0.2007 \times 0.045} = \sqrt{0.009032} = 0.095$$

$$Z = \frac{0.130 - 0}{0.095} = 1.37$$

The two-sided p-value for $Z = 1.37$ is $2 \times P(Z > 1.37) = 2 \times 0.085 = 0.17$.

Step 4: Conclusion. The p-value of about 0.17 is larger than 0.05. We fail to reject H_0: the data do not provide convincing evidence that the blood thinner affects 24-hour survival rates. That is not the same as showing the blood thinner has no effect. The observed difference — 13 percentage points in favor of the blood thinner — is not small; a study of only 90 patients simply cannot distinguish a difference of that size from chance variation.

12.6.4 Why pool for hypothesis tests but not for confidence intervals?

You may have noticed a subtle difference:

- **Hypothesis tests** use the **pooled proportion** $\hat{p}_{pool}$ in the standard error. This is because under H_0, we assume $p_1 = p_2$, so the best estimate of this common proportion comes from combining both samples.
- **Confidence intervals** use the **individual sample proportions** $\hat{p}_1$ and $\hat{p}_2$ separately. We are not assuming the proportions are equal – we are estimating how different they are.

This distinction applies only to proportions. When we turn to means in the next chapters, we will instead use separate sample standard deviations in both tests and confidence intervals.

12.7 Comparing simulation-based and normal-based methods

Throughout this course, we have developed two parallel approaches to inference:

Table 12.3: Comparison of simulation-based and normal approximation approaches.

Feature	Simulation-based	Normal approximation
Requires conditions?	Minimal (mainly independence)	Independence + success-failure
Small samples?	Works well	May be inaccurate
Computational cost	Higher (needs many simulations)	Lower (one formula)
Conceptual clarity	High (directly simulates the process)	Moderate (requires distribution theory)
Precision	Approximate (varies by number of simulations)	Approximate (varies by how well conditions hold)

12.7.1 An example comparison

Consider the opportunity cost study from the randomization tests chapter. The observed difference in proportions was 0.20 ($\hat{p}_T - \hat{p}_C$), and we wanted to test $H_0 : p_T - p_C = 0$.

Randomization test (the randomization tests chapter): We shuffled the group labels 1,000 times. The p-value was about 0.005 (5 of the 1,000 shuffles).

Normal approximation (this chapter): The pooled standard error formula gives $SE = 0.078$ (close to the standard deviation of the shuffled differences), so $Z = 0.20/0.078 = 2.56$ and the p-value is 0.005.

Both methods give essentially the same answer and the same conclusion. This is reassuring and typical when conditions are met.

12.7.2 When do they disagree?

The methods can give different answers when:

1. **The success-failure condition fails.** The normal approximation may undercount the probability in the tails, leading to p-values that are too small. The simulation approach remains valid.
2. **The sampling distribution is noticeably skewed.** The normal model is symmetric by definition, while the true sampling distribution of $\hat{p}$ near 0 or 1 is skewed.

The medical consultant example illustrates this: with only 6.2 expected complications, the normal model gives a p-value of about 0.085, while the simulation earlier in this chapter gave 0.132 – about 55% higher. The simulation-based p-value is more trustworthy in this case.

12.7.3 Practical recommendation

- **When conditions are met**: Both methods work. Use the normal approximation for its simplicity, or use simulation for its conceptual directness.
- **When conditions are borderline or not met**: Use simulation. The bootstrap and randomization approaches make fewer assumptions and are more reliable in these settings.
- **Always check conditions first.** The success-failure condition is a simple, quick check that helps you decide which approach to use.

12.8 Explore with StatLens

The Sampling Distribution Lab (proportion mode) builds the sampling distribution of $\hat{p}$ one sample at a time. Set $p = 0.60$ and $n = 50$, watch the distribution fill in, then drop p to 0.05 to see what happens when the success–failure condition fails.

Analytical tools:

- One-Proportion z test + CI, Two-Proportion z test + CI.

Simulation counterparts (for when conditions fail):

- Bootstrap CI for One Proportion, Bootstrap CI for Difference in Proportions.
- Randomization test for One Proportion, Randomization test for Two Proportions.

12.9 Chapter review

12.9.1 Summary

This chapter developed normal-based inference methods for proportions:

- **Success-failure condition.** The normal approximation for $\hat{p}$ requires at least 10 expected successes and 10 expected failures. For confidence intervals, check with $\hat{p}$; for hypothesis tests, check with p_0.
- **CI for one proportion**: $\hat{p} \pm z^* \times \sqrt{\hat{p}(1-\hat{p})/n}$. Use $\hat{p}$ in the SE.
- **Hypothesis test for one proportion**: $Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$. Use p_0 in the SE.

- **CI for difference of proportions**: $(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE$, where SE uses separate sample proportions.
- **Hypothesis test for difference of proportions**: $Z = \frac{\hat{p}_1 - \hat{p}_2}{SE}$, where SE uses the **pooled proportion** $\hat{p}_{pool}$.
- **Confidence level** is controlled by z^*: higher confidence means wider intervals.
- **When conditions fail** (especially the success-failure condition), use simulation-based methods from the randomization tests and bootstrap confidence intervals chapters instead.
- **Normal-based and simulation-based methods agree** when conditions are met and can give different answers when conditions are violated.

12.9.2 Key terms

Success-failure condition, standard error of $\hat{p}$, Z test for a proportion, pooled proportion, confidence interval for a proportion, confidence interval for a difference of proportions, margin of error, critical value (z^*), confidence level.

12.10 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Do aliens exist?** In May 2021, YouGov asked 4,839 adult Great Britain residents whether they think aliens exist, and if so, if they have or have not visited Earth. You want to evaluate if more than a quarter (25%) of Great Britain adults think aliens don't exist. In the survey 22% responded "I think they exist, and have visited Earth", 28% responded "I think they exist, but have not visited Earth", 29% responded "I don't think they exist", and 22% responded "Don't know". A friend of yours offers to help you with setting up the hypothesis test and comes up with the following hypotheses. Indicate any errors you see.

$H_0 : \hat{p} = 0.29 \qquad H_A : \hat{p} > 0.29$

2. **Married at 25.** A study suggests that the 25% of 25 year-olds have gotten married. You believe that this is incorrect and decide to collect your own sample for a hypothesis test. From a random sample of 776 25 year-olds, you find that 24% of them are married. A friend of yours offers to help you with setting up the hypothesis test and comes up with the following hypotheses. Indicate any errors you see.

$H_0 : \hat{p} = 0.24 \qquad H_A : \hat{p} \neq 0.24$

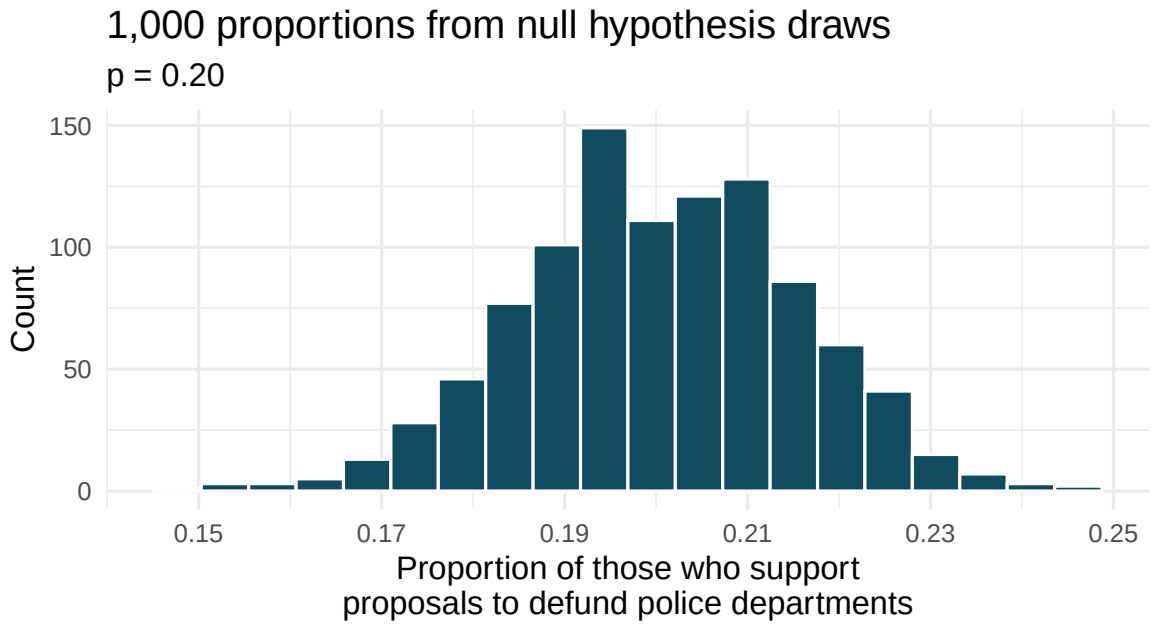
3. **Defund the police.** A Survey USA poll conducted in Seattle, WA in May 2021 reports that of the 650 respondents (adults living in this area), 159 support proposals to defund police departments. (Survey USA 2021)

a. A journalist writing a news story on the poll results wants to use the headline "More than 1 in 5 adults living in Seattle support proposals to defund police departments." You caution the journalist that they should first conduct a hypothesis test to see if the poll data provide convincing evidence for this claim Write the hypotheses for this test.

b. Calculate the proportion of Seattle adults in the sample who support proposals to defund police departments.

c. Describe a setup for a simulation that would be appropriate in this situation and how the p-value can be calculated using the simulation results.

d. The histogram below shows the distribution of 1,000 $\hat{p}_{sim}$s under the null hypothesis. Estimate the p-value using the plot and use it to evaluate the hypotheses.

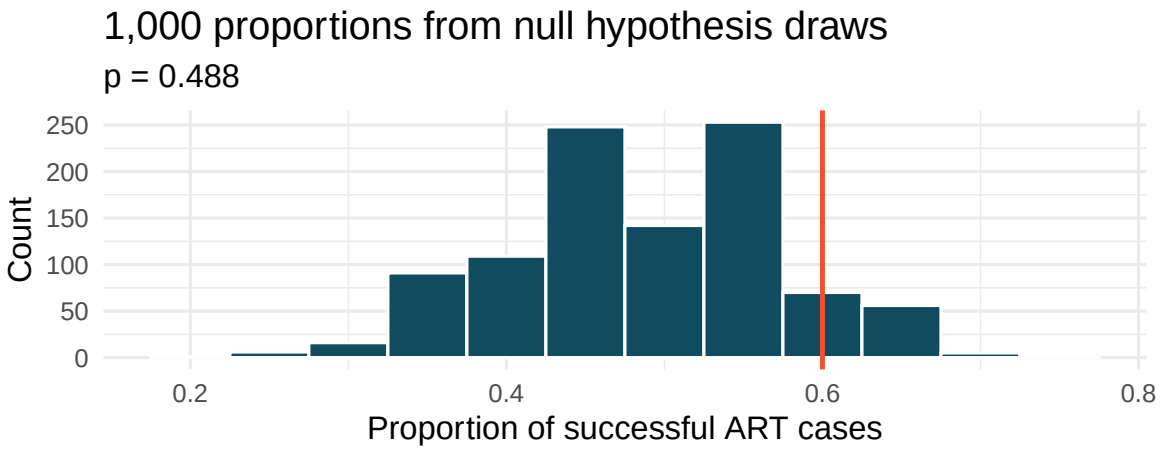

1,000 proportions from null hypothesis draws
p = 0.20
Count
150
100
50
0
0.15
0.17
0.19
0.21
0.23
0.25
Proportion of those who support
proposals to defund police departments

4. **Assisted reproduction.** Assisted Reproductive Technology (ART) is a collection of techniques that help facilitate pregnancy (e.g., in vitro fertilization). The 2018 ART Fertility Clinic Success Rates Report published by the Centers for Disease Control and Prevention reports that ART has been successful in leading to a live birth in 48.8% of cases where the patient is under 35 years old. (CDC 2018) A new fertility clinic claims that their success rate is higher than average for this age group. A random sample of 30 of their patients yielded a success rate of 60%. A consumer watchdog group would like to determine if the data provides strong evidence to support the company's claim.

a. Write the hypotheses to test if the success rate for ART at this clinic is discernibly higher than the success rate reported by the CDC.

b. Describe a setup for a simulation that would be appropriate in this situation and how the p-value can be calculated using the simulation results.

c. The histogram below shows the distribution of 1,000 $\hat{p}_{sim}$s under the null hypothesis. Estimate the p-value using the plot and use it to evaluate the hypotheses.

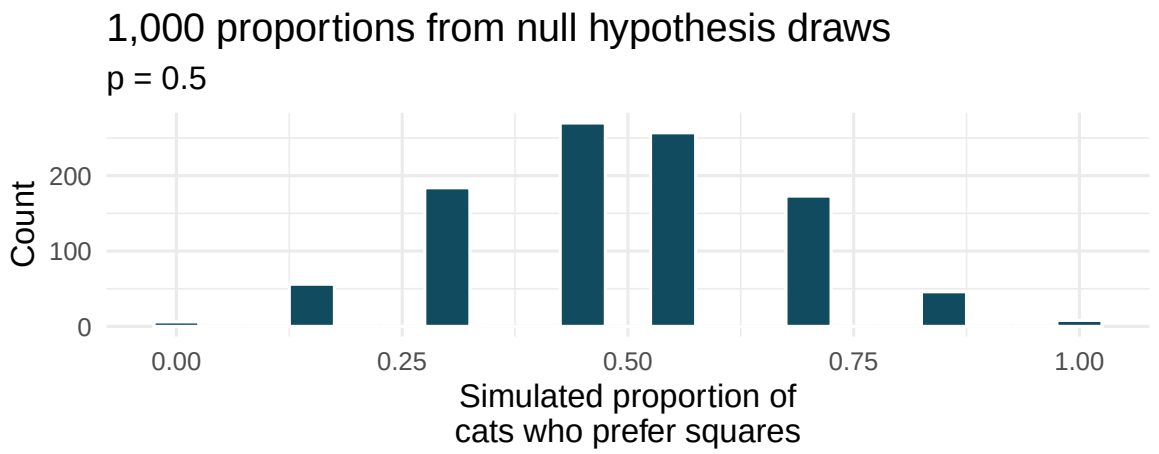


d. After performing this analysis, the consumer group releases the following news headline: "Infertility clinic falsely advertises better success rates". Comment on the appropriateness of this statement.

5. **If I fits, I sits, standard errors.** The results of a study on the type of enclosed spaces cats are most likely to sit in show that 5 out of 7 cats chose a square taped to the ground over a shape known as Kanizsa square illusion, which was preferred by the remaining 2 cats. To evaluate whether these data provide convincing evidence that cats prefer one of the shapes over the other, we set $H_0 : p = 0.5$, where p is the population proportion of cats who prefer square over the Kanizsa square illusion and $H_A : p \neq 0.5$, which suggests some preference, without specifying which shape is more preferred. (Smith, Chouinard, and Byosiere 2021)

a. Using the mathematical model, calculate the standard error of the sample proportion in repeated samples of size 7.

b. A null hypothesis simulation (with 1,000 draws) was run, and the resulting null distribution is displayed in the histogram below. This distribution shows the variability of the sample proportion in samples of size 7 when 50% of cats prefer the square shape over the Kanizsa square illusion. What is the approximate standard error of the sample proportion based on this distribution?



c. Do the mathematical model and simulated draws yield similar standard errors?

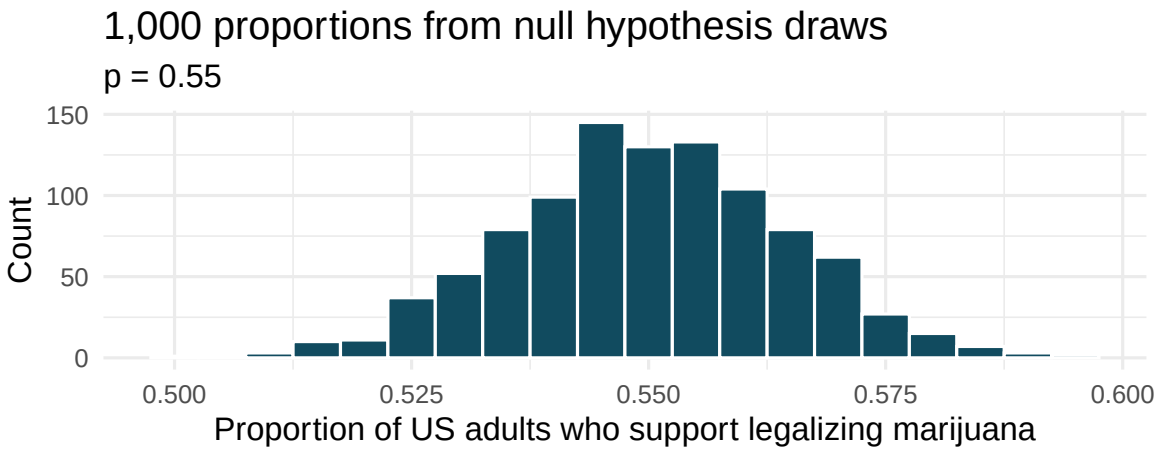
d. In order to approach the problem using the mathematical model, is the success-failure condition met for this study? Explain.

e. What features of the null distribution shown above tells us that the mathematical model should probably not be used?

6. **Legalization of marijuana, standard errors.** According to the 2022 General Social Survey, in a random sample of 1,207 US adults, 65.3% think marijuana should be made legal. (NORC 2022) Consider a scenario where, in order to become legal, 55% (or more) of voters must approve.

a. Calculate the standard error of the sample proportion using the mathematical model.

b. 1,000 sample proportions from samples of size 1,207 were drawn from a null distribution where 55% of voters approve legalizing marijuana. The distribution of these proportions is shown in the histogram below. Approximate the standard error of the sample proportion based on this distribution.

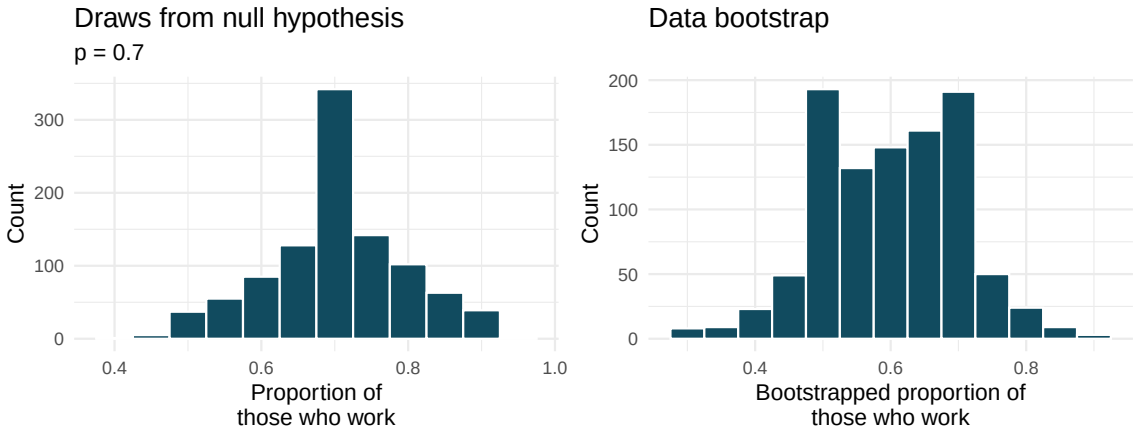


c. Do the mathematical model and simulated draws yield similar standard errors?

d. In this setting (to test whether the true underlying population proportion is greater than 0.55), would there be a strong reason to choose the mathematical model over the simulated null hypothesis (or vice versa)?

7. **Statistics and employment, use the bootstrap.** In a large university where 70% of the full-time students are employed at least 5 hours per week, the members of the Statistics Department wonder if the same proportion of their students work at least 5 hours per week. They randomly sample 25 majors and find that 15 of the students work 5 or more hours each week.

Two sampling distributions are created to describe the variability in the proportion of statistics majors who work at least 5 hours per week. The null hypothesis distribution imposes a true population proportion of $p = 0.7$ while the data bootstrap resamples from the actual data (which has 60% of the observations who work at least 5 hours per week).



See the next page for the questions.

a. Which distribution should be used to test whether the proportion of all statistics majors who work at least 5 hours per week is 70%? And which distribution should be used to find a confidence interval for the true poportion of statistics majors who work at least 5 hours per week?

b. Using the appropriate histogram, test the claim that 70% of statistics majors, like their peers, work at least 5 hours per week. State the null and alternative hypotheses, find the p-value, and conclude the test in the context of the problem.

c. Using the appropriate histogram, find a 98% bootstrap percentile confidence interval for the true proportion of statistics majors who work at least 5 hours per week. Interpret the confidence interval in the context of the problem.

d. Using the appropriate historgram, find a 98% bootstrap SE confidence interval for the true proportion of statistics majors who work at least 5 hours per week. Interpret the confidence interval in the context of the problem.

8. **CLT for proportions.** Define the term "sampling distribution" of the sample proportion, and describe how the shape, center, and spread of the sampling distribution change as the sample size increases when $p = 0.1$.

9. **Vegetarian college students.** Suppose that 8% of college students are vegetarians. Determine if the following statements are true or false, and explain your reasoning.

a. The distribution of the sample proportions of vegetarians in random samples of size 60 is approximately normal since $n \geq 30$.

b. The distribution of the sample proportions of vegetarian college students in random samples of size 50 is right skewed.

c. A random sample of 125 college students where 12% are vegetarians would be considered unusual.

d. A random sample of 250 college students where 12% are vegetarians would be considered unusual.

e. The standard error would be reduced by one-half if we increased the sample size from 125 to 250.

10. **Young Americans, American dream.** About 77% of young adults think they can achieve the American dream. Determine if the following statements are true or false, and explain your reasoning. (Vaughn 2011)

a. The distribution of sample proportions of young Americans who think they can achieve the American dream in random samples of size 20 is left skewed.

b. The distribution of sample proportions of young Americans who think they can achieve the American dream in random samples of size 40 is approximately normal since $n \geq 30$.

c. A random sample of 60 young Americans where 85% think they can achieve the American dream would be considered unusual.

d. A random sample of 120 young Americans where 85% think they can achieve the American dream would be considered unusual.

11. **Orange tabbies.** Suppose that 90% of orange tabby cats are male. Determine if the following statements are true or false, and explain your reasoning.

a. The distribution of sample proportions of random samples of size 30 is left skewed.

b. Using a sample size that is 4 times as large will reduce the standard error of the sample proportion by one-half.

c. The distribution of sample proportions of random samples of size 140 is approximately normal.

d. The distribution of sample proportions of random samples of size 280 is approximately normal.

12. **Young Americans, starting a family.** About 25% of young Americans have delayed starting a family due to the continued economic slump. Determine if the following statements are true or false, and explain your reasoning. (Demos 2011)

a. The distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump in random samples of size 12 is right skewed.

b. In order for the distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump to be approximately normal, we need random samples where the sample size is at least 40.

c. A random sample of 50 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

d. A random sample of 150 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

e. Tripling the sample size will reduce the standard error of the sample proportion by one-third.

13. **Sex equality.** The General Social Survey asked a random sample of 1,390 Americans the following question: "On the whole, do you think it should or should not be the government's responsibility to promote equality between men and women?" 82% of the respondents said it "should be". At a 95% confidence level, this sample has 2% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning. (NORC 2016)

a. We are 95% confident that 80% to 84% of Americans in this sample think it's the government's responsibility to promote equality between men and women.

b. We are 95% confident that 80% to 84% of all Americans think it's the government's responsibility to promote equality between men and women.

c. If we considered many random samples of 1,390 Americans, and we calculated 95% confidence intervals for each, 95% of these intervals would include the true population proportion of Americans who think it's the government's responsibility to promote equality between men and women.

d. In order to decrease the margin of error to 1%, we would need to quadruple (multiply by 4) the sample size.

e. Based on this confidence interval, there is sufficient evidence to conclude that a majority of Americans think it's the government's responsibility to promote equality between men and women.

14. **Elderly drivers.** A Marist Poll report states that 66% of American adults think licensed drivers should be required to retake their road test once they reach 65 years of age, based on a random sample of 1,018 American adults. They also report a margin of error was 3% at the 95% confidence level. (Poll 2011)

a. Verify the margin of error reported by The Marist Poll using a mathematical model.

b. Based on a 95% confidence interval, does the poll provide convincing evidence that *more than* two thirds of the population think that licensed drivers should be required to retake their road test once they turn 65?

15. **Fireworks on July 4th.** A local news outlet reported that 56% of 600 randomly sampled Kansas residents planned to set off fireworks on July 4th. Determine the margin of error for the 56% point estimate using a 95% confidence level using a mathematical model. (Survey USA 2012)

16. **Proof of COVID-19 vaccination.** In the US, businesses and schools shut down due to the COVID-19 pandemic in March 2020, and a vaccine became publicly available for the first time in April 2021. That month, a Gallup poll surveyed a random sample of 3,731 US adults, asking how they felt about the COVID-19 vaccine requirement for air travel. The poll found that 57% said they would favor it. (Gallup 2021b)

a. Describe the population parameter of interest. What is the value of the point estimate of this parameter?

b. Check if the conditions required for constructing a confidence interval using a mathematical model based on these data are met.

c. Construct a 95% confidence interval for the proportion of US adults who favor requiring proof of COVID-19 vaccination for travel by airplane.

d. Without doing any calculations, describe what would happen to the confidence interval if we decided to use a higher confidence level.

e. Without doing any calculations, describe what would happen to the confidence interval if we used a larger sample.

17. **Study abroad.** A survey on 1,509 high school seniors who took the SAT and who completed an optional web survey shows that 55% of high school seniors are fairly certain that they will participate in a study abroad program in college. (American Council on Education 2008)

a. Is this sample a representative sample from the population of all high school seniors in the US? Explain your reasoning.

b. Suppose the conditions for inference are met, regardless of your answer to part (a). Using a mathematical model, construct a 90% confidence interval for the proportion of high school seniors (of those who took the SAT) who are fairly certain they will participate in a study abroad program in college, and interpret this interval in context.

c. What does "90% confidence" mean?

d. Based on this interval, would it be appropriate to claim that the majority of high school seniors are fairly certain that they will participate in a study abroad program in college?

18. **Legalization of marijuana, mathematical interval.** The General Social Survey asked a random sample of 1,563 US adults: "Do you think the use of marijuana should be made legal, or not?" 60% of the respondents said it should be made legal. (NORC 2022)

a. Is 60% a sample statistic or a population parameter? Explain.

b. Using a mathematical model, construct a 95% confidence interval for the proportion of US adults who think marijuana should be made legal, and interpret it in the context of the data.

c. A critic points out that this 95% confidence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Do the technical conditions hold for these data? Explain.

d. A news piece on this survey's findings states, "Majority of US adults think marijuana should be legalized." Based on your confidence interval, is the news piece's statement justified?

19. **National Health Plan, mathematical inference.** A Kaiser Family Foundation poll for a random sample of US adults in 2019 found that 79% of Democrats, 55% of Independents, and 24% of Republicans supported a generic "National Health Plan". There were 347 Democrats, 298 Republicans, and 617 Independents surveyed. (Kaiser Family Foundation 2019)

    a. A political pundit on TV claims that a majority of Independents support a National Health Plan. Do these data provide strong evidence to support this type of statement? Your response should use a mathematical model.

    b. Would you expect a confidence interval for the proportion of Independents who oppose the public option plan to include 0.5? Explain.

20. **Is college worth it?** Among a simple random sample of 331 American adults who do not have a four-year college degree and are not currently enrolled in school, 48% said they decided not to go to college because they could not afford school. (Pew Research Center 2011)

    a. A newspaper article states that only a minority of the Americans who decide not to go to college do so because they cannot afford it and uses the point estimate from this survey as evidence. Conduct a hypothesis test to determine if these data provide strong evidence supporting this statement.

    b. Would you expect a confidence interval for the proportion of American adults who decide not to go to college because they cannot afford it to include 0.5? Explain.

21. **Taste test.** Some people claim that they can tell the difference between a diet soda and a regular soda in the first sip. A researcher wanting to test this claim randomly sampled 80 such people. He then filled 80 plain white cups with soda, half diet and half regular through random assignment, and asked each person to take one sip from their cup and identify the soda as diet or regular. 53 participants correctly identified the soda.

    a. Do these data provide strong evidence that these people are able to detect the difference between diet and regular soda, in other words, are the results discernibly better than just random guessing? Your response should use a mathematical model.

    b. Interpret the p-value in this context.

22. **Will the coronavirus bring the world closer together?** In early 2020 the COVID-19 pandemic arrived in the US; by December 2020 the first COVID-19 vaccine was available. An April 2021 YouGov poll asked 4,265 UK adults whether they think the coronavirus bring the world closer together or leave us further apart. 12% of the respondents said it will bring the world closer together. 37% said it would leave us further apart, 39% said it won't make a difference and the remainder didn't have an opinion on the matter. (YouGov 2021)

    a. Calculate, using a mathematical model, a 90% confidence interval for the proportion of UK adults who think the coronavirus will bring the world closer together, and interpret the interval in context.

    b. Suppose we wanted the margin of error for the 90% confidence level to be about 0.5%. How large of a sample size would you recommend for the poll?

23. **Quality control.** As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

    a. What population is under consideration in the dataset?

    b. What parameter is being estimated?

    c. What is the point estimate for the parameter?

    d. What is the name of the statistic that can be used to measure the uncertainty of the point estimate?

    e. Compute the value of the statistic from part (d) using a mathematical model.

    f. The historical rate of defects is 10%. Should the engineer be surprised by the observed rate of defects during the current week?

    g. Suppose the true population value was found to be 10%. If we use this proportion to recompute the value in part (d) using $p = 0.1$ instead of $\hat{p}$, how much does the resulting value of the statistic change?

24. **Nearsighted children.** Nearsightedness (myopia) is a common vision condition in which you can see near objects clearly, but farther away objects blurry. It is believed that nearsightedness affects about 8% of all children. In a random sample of 194 children, 21 are nearsighted. Using a mathematical model, conduct a hypothesis test for the following question: do these data provide evidence that the 8% value is inaccurate?

25. **Website registration.** A website is trying to increase registration for first-time visitors, exposing 1% of these visitors to a new site design. Of 752 randomly sampled visitors over a month who saw the new design, 64 registered.

    a. Check the conditions for constructing a confidence interval for the proportion of first-time visitors of the site who would register under the new design using a mathematical model.

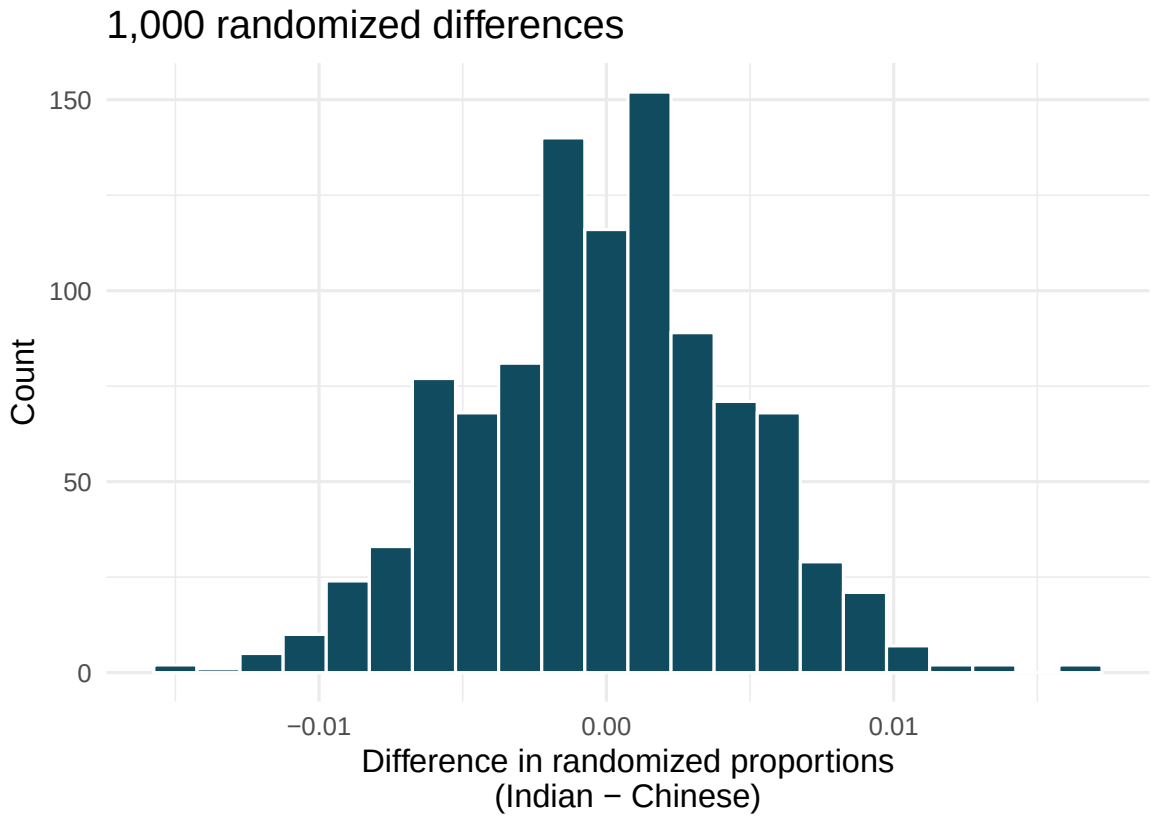
    b. Compute the standard error which would describe the variability os the point estimate associated with repeated samples of size 752.

    c. Construct and interpret a 90% confidence interval for the fraction of first-time visitors of the site who would register under the new design (assuming stable behaviors by new visitors over time).

26. **Coupons driving visits.** A store randomly samples 603 shoppers over the course of a year and finds that 142 of them made their visit because of a coupon they'd received in the mail. Using a mathematical model, construct a 95% confidence interval for the fraction of all shoppers during the year whose visit was because of a coupon they'd received in the mail.

27. **Disaggregating Asian American tobacco use, hypothesis testing.** Understanding cultural differences in tobacco use across different demographic groups can lead to improved health care education and treatment. A recent study disaggregated tobacco use across Asian American ethnic groups including Asian-Indian (n = 4,373), Chinese (n = 4,736), and Filipino (n = 4,912), in comparison to non-Hispanic Whites (n = 275,025). The number of current smokers in each group was reported as Asian-Indian (n = 223), Chinese (n = 279), Filipino (n = 609), and non-Hispanic Whites (n = 50,880). (Rao et al. 2021)

To determine whether the proportion of Asian-Indian Americans who are current smokers is different from the proportion of Chinese Americans who are smokers, a randomization simulation was performed.

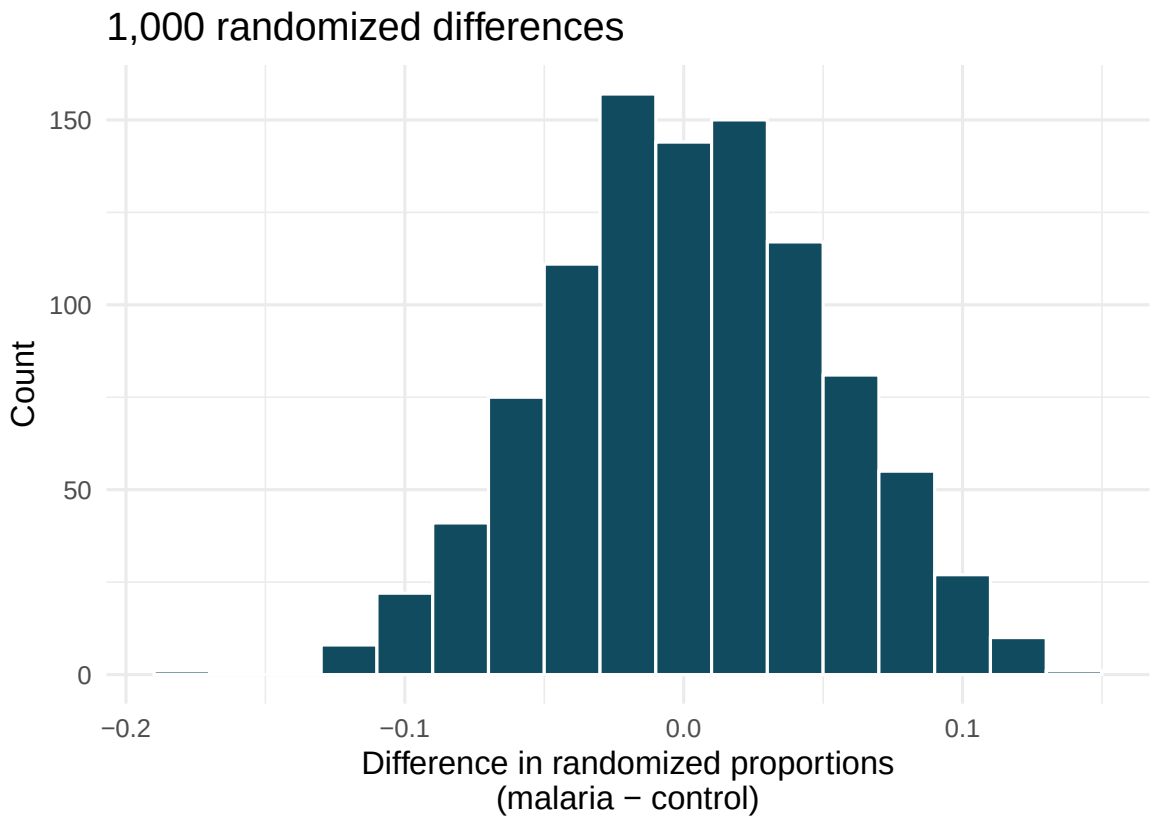


a. In both words and symbols provide the parameter and statistic of interest for this study. Do you know the numerical value of either the parameter or statistic of interest? If so, provide the numerical value.

b. The histogram above provides the sampling distribution (under randomization) for $\hat{p}_{Asian-Indian} - \hat{p}_{Chinese}$ under repeated null randomizations ($\hat{p}$ is the proportion in the sample who are current smokers). Estimate the standard error of $\hat{p}_{Asian-Indian} - \hat{p}_{Chinese}$ based on the randomization histogram.

c. Consider the hypothesis test to determine if there is a difference in proportion of Asian-Indian Americans as compared to Chinese Americans who are current smokers. Write out the null and alternative hypotheses, estimate a p-value using the randomization histogram, and conclude the test in the context of the problem.

28. **Malaria vaccine effectiveness, hypothesis test.** With no currently licensed vaccines to inhibit malaria, good news was welcomed with a recent study reporting long-awaited vaccine success for children in Burkina Faso. With 450 children randomized to either one of two different doses of the malaria vaccine or a control vaccine, 89 of 292 malaria vaccine and 106 out of 147 control vaccine children contracted malaria within 12 months after the treatment. (Datoo et al. 2021)

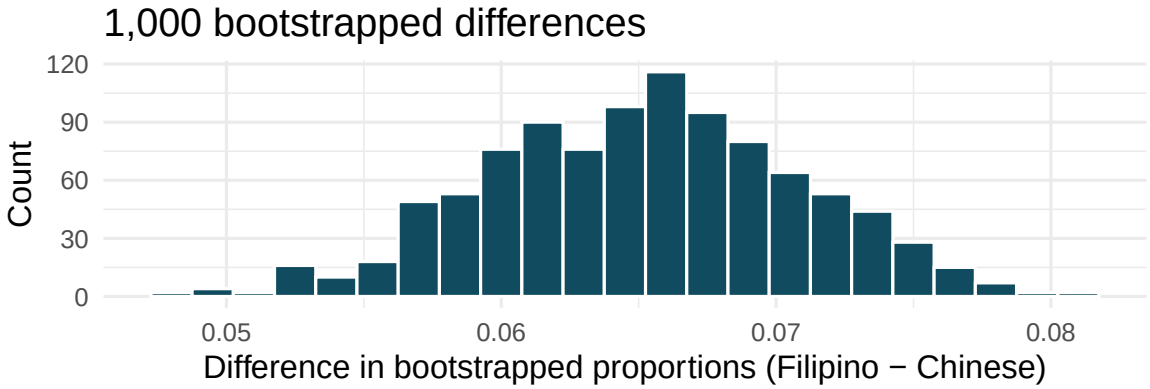


a. In both words and symbols provide the parameter and statistic of interest for this study. Do you know the numerical value of either the parameter or statisic of interest? If so, provide the numerical value.

b. The histogram above provides the sampling distribution (under randomization) for $\hat{p}_{malaria} - \hat{p}_{control}$ under repeated null randomizations ($\hat{p}$ is the proportion of children in the sample who contracted malaria). Estimate the standard error of $\hat{p}_{malaria} - \hat{p}_{control}$ based on the randomization histogram.

c. Consider the hypothesis test constructed to show a lower proportion of children contracting malaria on the malaria vaccine as compared to the control vaccine. Write out the null and alternative hypotheses, estimate a p-value using the randomization histogram, and conclude the test in the context of the problem.

29. **Disaggregating Asian American tobacco use, confidence interval.** Based on a study on the degree to which smoking practices differ across ethnic groups. a confidence interval for the difference in current smoking status for Filipino versus Chinese Americans is desired. (Rao et al. 2021)

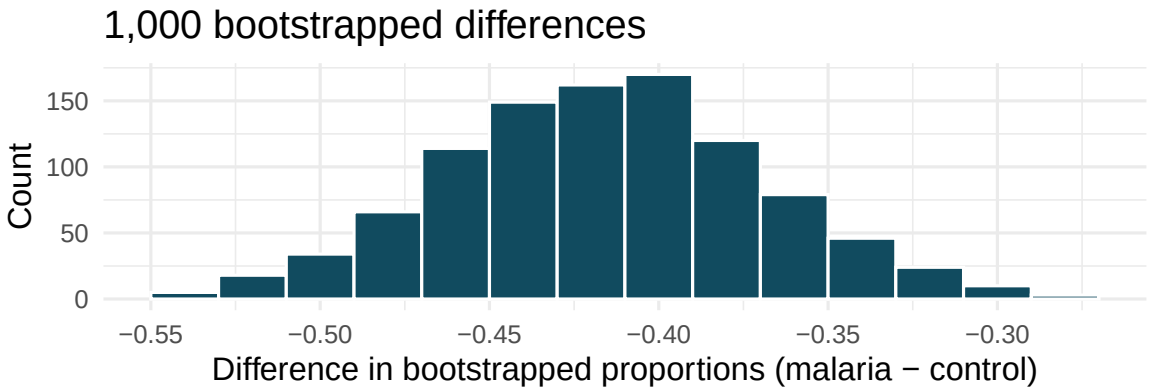


a. Consider the bootstrap distribution of difference in sample proportions of current smokers (Filipino Americans minus Chinese Americans) in 1,000 bootstrap repetitions as above. Estimate the standard error of the difference in sample proportions, as seen in the histogram.

b. Using the standard error from the bootstrap distribution, find a 95% bootstrap SE confidence interval for the true difference in proportion of current smokers (Filipino Americans minus Chinese Americans) in the population. Interpret the interval in the context of the problem.

c. Using the entire bootstrap distribution, find a 95% bootstrap percentile confidence interval for the true difference in proportion of current smokers (Filipino Americans minus Chinese Americans) in the population. Interpret the interval in the context of the problem.

30. **Malaria vaccine effectiveness, confidence interval.** With no currently licensed vaccines to inhibit malaria, good news was welcomed with a recent study reporting long-awaited vaccine success for children in Burkina Faso. With 450 children randomized to either one of two different doses of the malaria vaccine or a control vaccine, 89 of 292 malaria vaccine and 106 out of 147 control vaccine children contracted malaria within 12 months after the treatment. (Datoo et al. 2021)



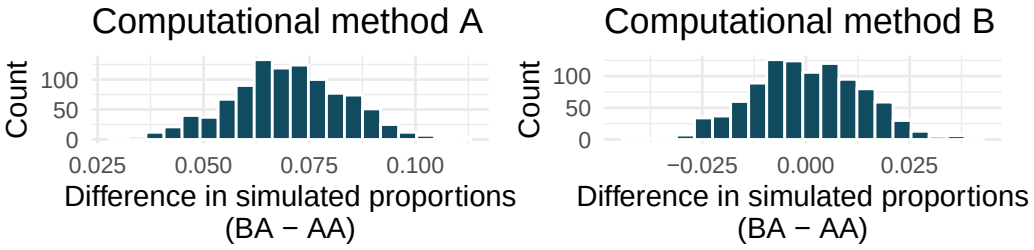
a. Consider the bootstrap distribution of difference in sample proportions of children who contracted malaria (malaria vaccine minus control vaccine) in 1000 bootstrap repetitions as above. Estimate the standard error of the difference in sample proportions, as seen in the histogram.

b. Using the standard error from the bootstrap distribution, find a 95% bootstrap SE confidence interval for the true difference in proportion of children who contract malaria (malaria vaccine minus control vaccine) in the population. Interpret the interval in the context of the problem.

c. Using the entire bootstrap distribution, find a 95% bootstrap percentile confidence interval for the true difference in proportion of children who contract malaria (malaria vaccine minus control vaccine) in the population. Interpret the interval in the context of the problem.

31. **COVID-19 and degree completion.** A 2021 Gallup poll surveyed 3,941 students pursuing a bachelor's degree and 2,064 students pursuing an associate degree (students were not randomly selected but were weighted so as to represent a random selection of currently enrolled US college students). The poll found that 51% of the bachelor's degree students and 44% of associate degree students said that the COVID-19 pandemic will negatively impact their ability to complete the degree. (Gallup 2021a)

Below are two histograms generated with different computational approaches (both use 1,000 repetitions) to research questions which could be asked of these data. One of the histograms can be used to do a randomization test on whether the proportions of bachelor's and associate students who think the COVID-19 pandemic will negatively impact their ability to complete the degree. The other histogram is a bootstrap distribution used to quantify the difference in the proportions of bachelor's and associate's students who feel this way.



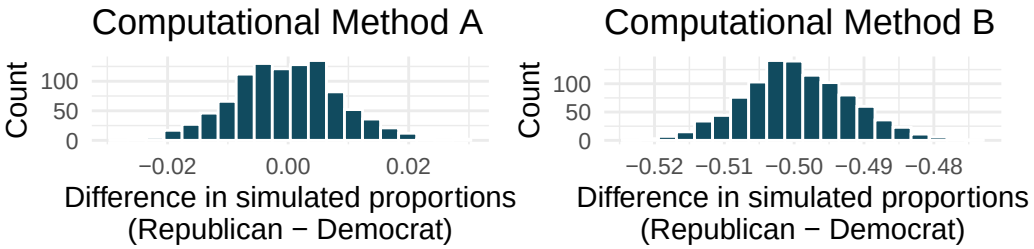
a. Are the center and standard error of the two graphs approximately the same? Explain.

b. Write a research question that can be addressed using the histogram generated with computational method A.

c. Write a research question that can addressed using the histogram generated with computational method B.

32. **Renewable energy.** A 2021 Gallup poll surveyed 5,447 randomly sampled US adults who are Republican (or Republican leaning) and 7,962 who are Democrats (or Democrat leaning). 31% of Republicans and 81% of Democrats said "government regulations are necessary to encourage businesses and consumers to rely more on renewable energy sources". (Gallup 2021a)

Below are two histograms generated with different computational approaches (both use 1,000 repetitions) to research questions which could be asked of these data. One of the histograms can be used to do a randomization test on whether the proportions of Republicans and Democrats who think government regulations are necessary to encourage businesses and consumers to rely more on renewable energy sources are different. The other histogram is a bootstrap distribution used to quantify the difference in the proportions of Republicans and Democrats who agree with this statement.



a. Are the center and standard error of the two graphs approximately the same? Explain.

b. Write a research question that can addressed using the histogram generated with computational method A.

c. Write a research question that can addressed using the histogram generated with computational method B.

33. **HIV in sub-Saharan Africa.** In July 2008 the US National Institutes of Health announced that it was stopping a clinical study early because of unexpected results. The study population consisted of HIV-infected women in sub-Saharan Africa who had been given single dose Nevaripine (a treatment for HIV) while giving birth, to prevent transmission of HIV to the infant. The study was a randomized comparison of continued treatment of a woman (after successful childbirth) with Nevaripine vs Lopinavir, a second drug used to treat HIV. 240 women participated in the study; 120 were randomized to each of the two treatments. Twenty-four weeks after starting the study treatment, each woman was tested to determine if the HIV infection was becoming worse (an outcome called *virologic failure*). Twenty-six of the 120 women treated with Nevaripine experienced virologic failure, while 10 of the 120 women treated with the other drug experienced virologic failure. (Lockman et al. 2007)

a. Create a two-way table presenting the results of this study.

b. State appropriate hypotheses to test for difference in virologic failure rates between treatment groups.

c. Complete the hypothesis test and state an appropriate conclusion. (Reminder: Verify any necessary conditions for the test.)

34. **Supercommuters.** The fraction of workers who are considered "supercommuters", because they commute more than 90 minutes to get to work, varies by state. Suppose the 1% of Nebraska residents and 6% of New York residents are supercommuters. Now suppose that we plan a study to survey 1000 people from each state, and we will compute the sample proportions $\hat{p}_{NE}$ for Nebraska and $\hat{p}_{NY}$ for New York.

a. What is the associated mean and standard deviation of $\hat{p}_{NE}$ in repeated samples of size 1000?

b. What is the associated mean and standard deviation of $\hat{p}_{NY}$ in repeated samples of size 1000?

c. Calculate and interpret the mean and standard deviation associated with the difference in sample proportions for the two groups, $\hat{p}_{NY} - \hat{p}_{NE}$ in repeated samples of 1000 in each group.

d. How are the standard deviations from parts (a), (b), and (c) related?

35. **National Health Plan.** A Kaiser Family Foundation poll for US adults in 2019 found that 79% of Democrats, 55% of Independents, and 24% of Republicans supported a generic "National Health Plan". There were 347 Democrats, 298 Republicans, and 617 Independents surveyed. 79% of 347 Democrats and 55% of 617 Independents support a National Health Plan. (Kaiser Family Foundation 2019)

a. Calculate a 95% confidence interval for the difference between the proportion of Democrats and Independents who support a National Health Plan $(p_D - p_I)$, and interpret it in this context. We have already checked conditions for you.

b. True or false: If we had picked a random Democrat and a random Independent at the time of this poll, it is more likely that the Democrat would support the National Health Plan than the Independent.

36. **Sleep deprivation, CA vs. OR, confidence interval.** According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data. (CDC 2008)

37. **Gender pay gap in medicine.** A study examined the average pay for men and women entering the workforce as doctors for 21 different positions. (Lo Sasso et al. 2011)

a. If each gender was equally paid, then we would expect about half of those positions to have men paid more than women and women would be paid more than men in the other half of positions. Write appropriate hypotheses to test this scenario.

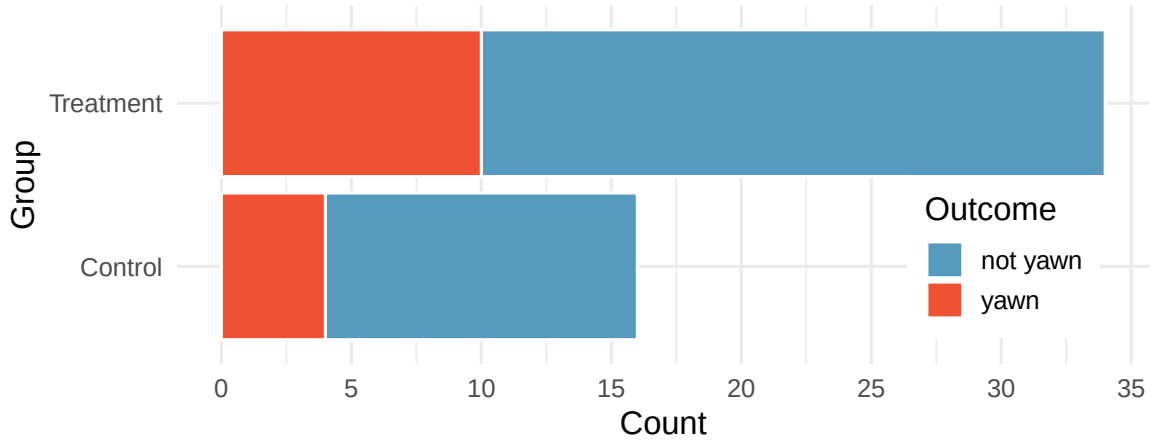
b. Men were, on average, paid more in 19 of those 21 positions. Complete a hypothesis test using your hypotheses from part (a).

38. **Sleep deprivation, CA vs. OR, hypothesis test.** A CDC report on sleep deprivation rates shows that the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents.

a. Conduct a hypothesis test to determine if these data provide strong evidence that the rate of sleep deprivation is different for the two states. (Reminder: Check conditions)

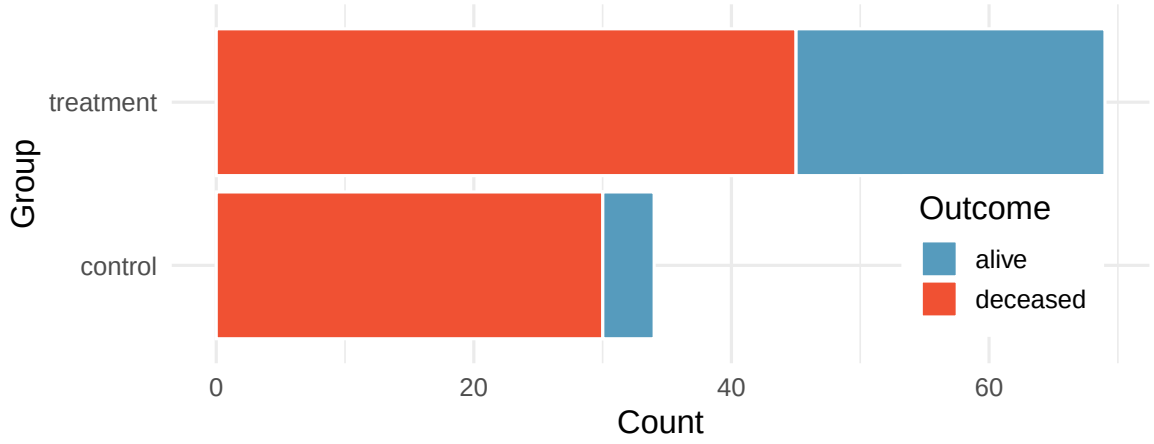
b. It is possible the conclusion of the test in part (a) is incorrect. If this is the case, what type of error was made?

39. **Is yawning contagious?** An experiment conducted by the MythBusters, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously influenced into yawning if another person near them yawns. 50 people were randomly assigned to two groups: 34 to a group where a person near them yawned (treatment) and 16 to a group where there wasn't a person yawning near them (control). The visualization below displays how many participants yawned in each group.



Suppose we are interested in estimating the difference in yawning rates between the control and treatment groups using a confidence interval. Explain why we cannot construct such an interval using the normal approximation. What might go wrong if we constructed the confidence interval despite this problem?

40. **Heart transplant success.** The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was officially designated a heart transplant candidate, meaning that he was gravely ill and might benefit from a new heart. Patients were randomly assigned into treatment and control groups. Patients in the treatment group received a transplant, and those in the control group did not. The visualization below displays how many patients survived and died in each group. (Turnbull, Brown, and Hu 1974)



Suppose we are interested in estimating the difference in survival rate between the control and treatment groups using a confidence interval. Explain why we cannot construct such an interval using the normal approximation. What might go wrong if we constructed the confidence interval despite this problem?

41. **Government shutdown.** The United States federal government shutdown of 2018–2019 occurred from December 22, 2018 until January 25, 2019, a span of 35 days. A Survey USA poll of 614 randomly sampled Americans during this time period reported that 48% of those who make less than $40,000 per year and 55% of those who make $40,000 or more per year said the government shutdown has not at all affected them personally. A 95% confidence interval for $(p_{<40K} - p_{\geq 40K})$, where p is the proportion of those who said the government shutdown has not at all affected them personally, is (-0.16, 0.02). Based on this information, determine if the following statements are true or false, and explain your reasoning if you identify the statement as false. (Survey USA 2019)

a. At the 5% discernibility level, the data provide convincing evidence of a real difference in the proportion who are not affected personally between Americans who make less than $40,000 annually and Americans who make $40,000 annually.

b. We are 95% confident that 16% more to 2% fewer Americans who make less than $40,000 per year are not at all personally affected by the government shutdown compared to those who make $40,000 or more per year.

c. A 90% confidence interval for $(p_{<40K} - p_{\geq 40K})$ would be wider than the $(-0.16, 0.02)$ interval.

d. A 95% confidence interval for $(p_{\geq 40K} - p_{<40K})$ is (-0.02, 0.16).

42. **Online harassment.** A Pew Research poll asked US adults aged 18-29 and 30-49 whether they have personally experienced harassment online. A 95% confidence interval for the difference between the proportions of 18-29 year-olds and 30-49 year-olds who have personally experienced harassment online $(p_{18-29} - p_{30-49})$ was calculated to be (0.115, 0.185). Based on this information, determine if the following statements are true or false, and explain your reasoning for each statement you identify as false. (Pew Research Center 2021b)

a. We are 95% confident that the true proportion of 18-29 year-olds who have personally experienced harassment online is 11.5% to 18.5% lower than the true proportion of 30-49 year-olds who have personally experienced harassment online.

b. We are 95% confident that the true proportion of 18-29 year-olds who have personally experienced harassment online is 11.5% to 18.5% higher than the true proportion of 30-49 year-olds who have personally experienced harassment online.

c. 95% of random samples will produce 95% confidence intervals that include the true difference between the population proportions of 18-29 year-olds and 30-49 year-olds who have personally experienced harassment online.

d. We can conclude that there is a discernible difference between the proportions of 18-29 year-olds and 30-49 year-olds who have personally experienced harassment online is too large to plausibly be due to chance, if in fact there is no difference between the two proportions.

e. The 90% confidence interval for $(p_{18-29} - p_{30-49})$ cannot be calculated with only the information given in this exercise.

43. **Decision errors and comparing proportions I.** In the following research studies, conclusions were made based on the data provided. It is always possible that the analysis conclusion could be wrong, although we will almost never actually know if an error has been made or not. For each study conclusion, specify which of a Type I or Type II error could have been made, and state the error in the context of the problem.

a. The malaria vaccine was seen to be effective at lowering the rate of contracting malaria (when compared to the control vaccine).

b. In the US population, Asian-Indian Americans and Chinese Americans are not observed to have different proportions of current smokers.

c. There is no evidence to claim a difference in the proportion of Americans who are not affected personally by a government shutdown when comparing Americans who make less than $40,000 annually and Americans who make $40,000 annually.

44. **Decision errors and comparing proportions II.** In the following research studies, conclusions were made based on the data provided. It is always possible that the analysis conclusion could be wrong, although we will almost never actually know if an error has been made or not. For each study conclusion, specify which of a Type I or Type II error could have been made, and state the error in the context of the problem.

a. Of registered voters in California, the proportion who report not knowing enough to voice an opinion on whether they support off shore drilling is different across those who have a college degree and those who do not.

b. In comparing Californians and Oregonians, there is no evidence to support a difference in the proportion of each who are sleep deprived.

45. **Active learning.** A teacher wanting to increase the active learning component of her course is concerned about student reactions to changes she is planning to make. She conducts a survey in her class, asking students whether they believe more active learning in the classroom (hands on exercises) instead of traditional lecture will helps improve their learning. She does this at the beginning and end of the semester and wants to evaluate whether students' opinions have changed over the semester. Can she used the methods we learned in this chapter for this analysis? Explain your reasoning.

46. **An apple a day keeps the doctor away.** A physical education teacher at a high school wanting to increase awareness on issues of nutrition and health asked her students at the beginning of the semester whether they believed the expression "an apple a day keeps the doctor away". 40% of the students responded yes. Throughout the semester she started each class with a discussion of a study highlighting positive effects of eating more fruits and vegetables. She conducted the same apple-a-day survey at the end of the semester, and this time 60% of the students responded yes. Can she used a two-proportion method from this section for this analysis? Explain your reasoning.

47. **Malaria vaccine effectiveness, effect size.** A randomized controlled trial on malaria vaccine effectiveness randomly assigned 450 children intro either one of two different doses of the malaria vaccine or a control vaccine. 89 of 292 malaria vaccine and 106 out of 147 control vaccine children contracted malaria within 12 months after the treatment. (Datoo et al. 2021)

Recall that in order to reject the null hypothesis that the two vaccines (malaria and control) are equivalent, we'd need the sample proportion to be about 2 standard errors below the hypothesized value of zero.

Say that the true difference (in the population) is given as δ, the sample sizes are the same in both groups ($n_{malaria} = n_{control}$), and the true proportion who contract malaria on the control vaccine is $p_{control} = 0.7$. If you ran your own study (in the future), how likely is it that you would get a difference in sample proportions that was sufficiently far from zero that you could reject under each of the conditions below. (*Hint:* Use the mathematical model.)

a. $\delta = -0.1$ and $n_{malaria} = n_{control} = 20$

b. $\delta = -0.4$ and $n_{malaria} = n_{control} = 20$

c. $\delta = -0.1$ and $n_{malaria} = n_{control} = 100$

d. $\delta = -0.4$ and $n_{malaria} = n_{control} = 100$

e. What can you conclude about values of δ and the sample size?

48. **Diabetes and unemployment.** A Gallup poll surveyed Americans about their employment status and whether they have diabetes. The survey results indicate that 1.5% of the 47,774 employed (full or part time) and 2.5% of the 5,855 unemployed 18-29 year-olds have diabetes. (Gallup 2012)

a. Create a two-way table presenting the results of this study.

b. State appropriate hypotheses to test for difference in proportions of diabetes between employed and unemployed Americans.

c. The sample difference is about 1%. If we completed the hypothesis test, we would find that the p-value is very small (about 0), meaning the difference is statistically discernible. Use this result to explain the difference between statistically discernible and practically important findings.

Datasets: `yawn` (openintro) | `heart_transplant` (openintro)

StatLens Exercises

These exercises focus on the parts of inference a calculator *can't* do for you: **framing the question**, **choosing the right procedure**, **checking conditions**, and **interpreting** what a tool reports. Where a link is given, let StatLens do the arithmetic and spend your effort on the reasoning.

1. **Frame the question.** A campus newspaper claims that *more than* 60% of students skip breakfast. A pollster surveys a random sample of 150 students and finds 99 who skip breakfast.
   a. What is the parameter of interest? Describe it in words and give its symbol.
   b. Write the null and alternative hypotheses in symbols. Is this a one-sided or two-sided test, and why?
   c. Check the success–failure condition for using a normal-approximation z-procedure here.

2. **Choose the procedure.** For each scenario, name the procedure you would use — **one-proportion** z (test or interval), **two-proportion** z (test or interval), or **not a proportion problem**. You do *not* need to carry anything out; the skill here is matching the question to the method.
   a. Is the proportion of defective parts from one assembly line different from the company's 2% target?
   b. Do a higher fraction of patients recover on a new drug than on the standard drug?
   c. Is the *average* recovery time on the new drug shorter than on the standard drug?
   d. In a single survey, estimate the percentage of voters who support a ballot measure, with a margin of error.

3. **Check conditions before you trust the test.** A medical consultant says her surgery patients have a *lower* complication rate than the national rate of 10%. Open the Medical Consultant data in the One-Proportion z-Test tool.
   a. State H_0 and H_A for the consultant's claim.
   b. The tool reports the sample size and the number of complications. Using $p_0 = 0.10$, compute np_0 and $n(1 - p_0)$. Is the success–failure condition met?
   c. Based on (b), is the normal-approximation z-test trustworthy here? Connect your answer to why we used a **bootstrap** confidence interval for this same dataset earlier in the course.

4. **Run it and interpret it (two proportions).** A randomized trial tested whether a malaria vaccine changes the infection rate. Open the Malaria vaccine trial data in the Two-Proportion z-Test tool.
   a. Identify the explanatory and response variables, and define the parameter $p_1 - p_2$ in words.
   b. State the hypotheses for testing whether the vaccine changes the infection rate.
   c. Run the test. Report the 95% confidence interval for $p_1 - p_2$ that the tool gives. Does the interval contain 0?
   d. Write a one-sentence conclusion **in context**, and say whether your confidence interval and your hypothesis test lead to the same decision.
   e. Which statement is the correct interpretation of the interval in (c)?
      i. There is a 95% probability that $p_1 - p_2$ lies in this interval.
      ii. If the trial were repeated many times, about 95% of the intervals built this way would contain the true difference $p_1 - p_2$.
      iii. 95% of vaccinated people fall within this interval.

5. **Simulation vs. analytic — when do they agree?** The normal-approximation z-procedure and a simulation (bootstrap) are two routes to the *same* confidence interval. Sometimes they land

in nearly the same place; sometimes they don't. You will compare them on two datasets and figure out *why*.

First the well-behaved case. Open the Stent study data in the **analytic** One-Proportion z tool, then open the *same* dataset in the bootstrap tool.

a. Record the 95% confidence interval from each tool. How close are the two intervals?

Now the troublesome case from Exercise 3. Open the Medical Consultant data in the analytic tool, then the same data in the bootstrap tool.

b. Record the 95% interval from each tool for this dataset. Are they as close as they were for the stent data? In particular, look at the *shape*: is the bootstrap distribution roughly symmetric, or is it skewed?
c. Explain the pattern. Why do the analytic and simulation intervals nearly coincide for the stent data but diverge for the medical-consultant data? Use the success–failure counts ($n\hat{p}$ and $n(1-\hat{p})$) from each dataset in your answer.
d. A student concludes, "Simulation and the formula disagree, so one of them must be wrong." Why is this the wrong takeaway? Which method should you *trust more* when they disagree, and why?

Chapter 13

Confidence Intervals for Means

In the bootstrap confidence intervals chapter, we used bootstrapping to construct confidence intervals – a flexible, simulation-based approach that works well in many settings. In the normal approximation chapter, we saw that sampling distributions are often well approximated by the normal distribution. For means, however, we encounter a complication: the population standard deviation σ is almost never known, and estimating it with the sample standard deviation s introduces extra uncertainty. To account for this, we use a slightly different bell-shaped distribution called the **t-distribution**. In this chapter, we develop t-based confidence intervals for one mean, the difference between two independent means, and the mean of paired differences.

13.1 The t-distribution

13.1.1 Why not the normal distribution?

Earlier we used the normal distribution to model sampling distributions, and in the previous chapter we applied it to proportions, where the standard error formula $SE = \sqrt{\hat{p}(1-\hat{p})/n}$ depends only on the sample proportion and the sample size – both of which are known. For means, the standard error is $SE = \sigma/\sqrt{n}$, which depends on σ, the population standard deviation. Since we rarely know σ, we estimate it using the sample standard deviation s:

$$SE \approx \frac{s}{\sqrt{n}}$$

This substitution introduces additional uncertainty, especially in small samples. When n is small, s can be a poor estimate of σ, and the normal model underestimates how variable the statistic really is. The t-distribution corrects for this.

13.1.2 Properties of the t-distribution

The **t-distribution** is a bell-shaped, symmetric distribution that is similar to the normal distribution but has **heavier tails** (more probability in the extremes). The exact shape of the t-distribution depends on a parameter called the **degrees of freedom** (**df**).

- When df is small, the t-distribution has noticeably heavier tails than the normal – extreme values are more likely.
- As df increases, the t-distribution approaches the standard normal distribution.
- For $df \geq 30$, the t-distribution is nearly indistinguishable from the normal.

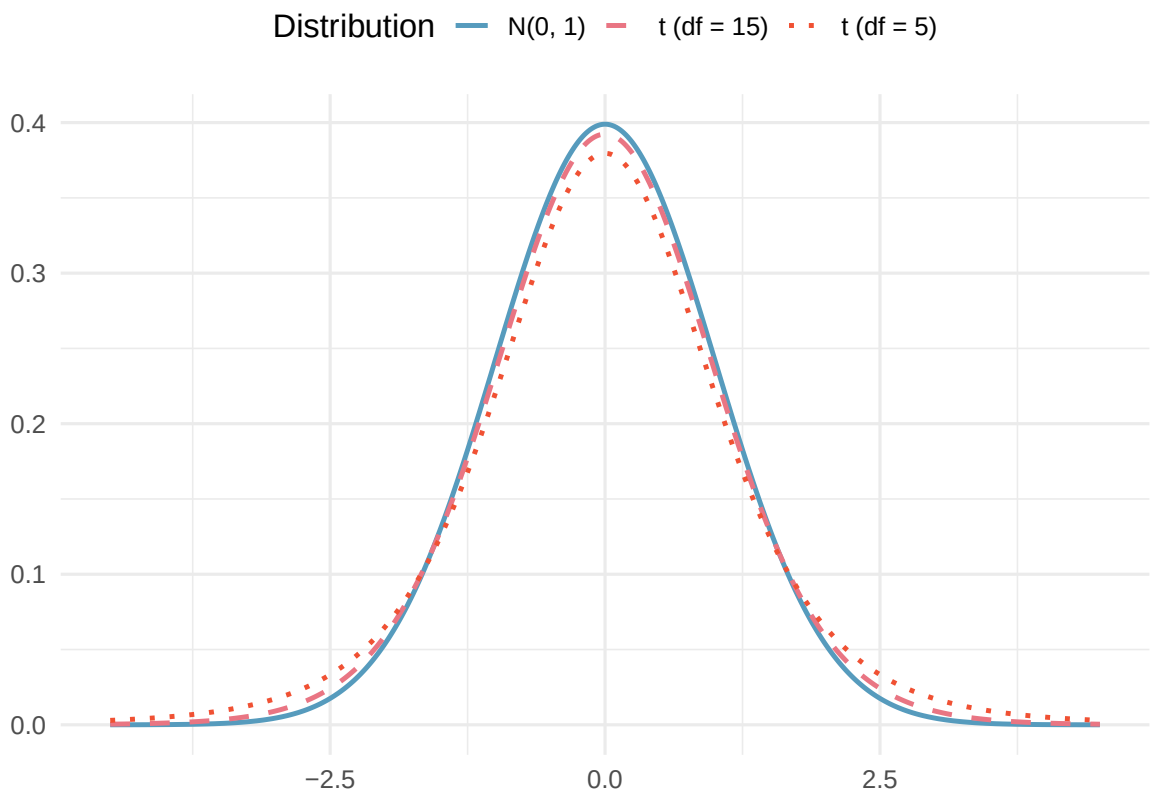


Figure 13.1: Three distributions plotted on the same axes: the standard normal distribution N(0,1), a t-distribution with df = 5, and a t-distribution with df = 15. All are centered at 0 and bell-shaped, but the t-distributions have heavier tails, especially when df is small. ↗ Try this live — drive the df slider yourself and watch the tails shrink as df grows.

Why heavier tails matter. Heavier tails mean that extreme values (large t-scores) are more probable under the t-distribution than under the normal distribution. In practice, this leads to **wider confidence intervals** when sample sizes are small – exactly the extra caution we need when σ is estimated by s.

How does the t-distribution compare to the normal distribution when $df = 2$? When $df = 100$?

Show answer

When $df = 2$, the t-distribution has very heavy tails – values far from 0 are much more likely than under the normal curve. When $df = 100$, the t-distribution is nearly identical to the standard normal. In general, the t-distribution becomes more like the normal as degrees of freedom increase.

See the t-distribution come alive. Open the t Distribution Explorer and drag the degrees of freedom slider. Start at $df = 2$ and watch the heavy tails — values far from zero are surprisingly common. Now slowly increase toward $df = 30$, then $df = 100$. The t-distribution gradually becomes indistinguishable from the normal curve. This is much harder to appreciate from a static figure — try it yourself.

13.1.3 Critical values from the t-distribution

Just as we use z^* for the normal distribution, we use t^*_{df} for the t-distribution. Because the t-distribution has heavier tails, t^* values are larger than the corresponding z^* values (especially for

small df).

Selected critical values for 95% confidence intervals:

Table 13.1: Critical values t^*_{df} for a 95% confidence interval. As df increases, t^* approaches $z^* = 1.96$.

df	t^*_{df} (95% CI)	z^* (for comparison)
5	2.571	1.960
10	2.228	1.960
20	2.086	1.960
30	2.042	1.960
50	2.009	1.960
100	1.984	1.960
∞	1.960	1.960

Why is the critical value $t^*_5 = 2.571$ so much larger than $z^* = 1.96$?

Show answer

With only $df = 5$, there is substantial uncertainty in estimating σ with s. The t-distribution accounts for this by having heavier tails, which pushes the critical values further from zero. This results in a wider confidence interval, reflecting the greater uncertainty.

13.2 One-sample t confidence interval

13.2.1 The setup

We want to estimate the population mean μ based on a single sample. The point estimate is the sample mean $\bar{x}$, and the standard error is $SE = s/\sqrt{n}$.

One-sample t confidence interval for μ.

$$\bar{x} \pm t^*_{df} \times \frac{s}{\sqrt{n}}$$

where $df = n - 1$.

Conditions:

1. **Independence.** The observations are independent (e.g., from a random sample).
2. **Normality / sample size.** Either the population distribution is approximately normal, or the sample size is large enough ($n \geq 30$). For smaller samples, check for strong skewness or extreme outliers.

13.2.2 Checking conditions

The normality condition is about the shape of the **population**, not the sample. In practice:

- If $n < 30$: Check a histogram or dotplot of the sample data. If there is no strong skewness and no extreme outliers, the t-methods are reasonable.
- If $n \geq 30$: The Central Limit Theorem ensures that $\bar{x}$ is approximately normal, even if the population is moderately skewed. However, be cautious with extreme outliers, which can distort $\bar{x}$ and s.

Outliers and the t-test. The sample mean $\bar{x}$ and sample standard deviation s are sensitive to outliers. A single extreme observation can substantially shift $\bar{x}$ and inflate s, leading to misleading confidence intervals. When outliers are present, consider whether they represent data errors. If the outliers are genuine, a bootstrap confidence interval (the bootstrap confidence intervals chapter) may be more reliable.

When neither t nor the plain bootstrap is enough. The t-interval assumes a moderately-behaved population (or a large sample); the percentile bootstrap fills in when normality is doubtful and n is at least moderate. But if your sample is **both small and strongly skewed**, or if extreme outliers dominate, *neither* tool is fully trustworthy. In that situation:

- **Increase the sample size** if you can — by far the best fix.
- **Use a rank-based method** (Wilcoxon signed-rank, sign test) that does not require estimating a mean.
- **Reach for the BCa bootstrap** — an accuracy-corrected CI that handles skew and small samples better than the plain percentile method. Available in StatLens under **expert mode** and in R via `boot::boot.ci(..., type = "bca")`.
- **Or consult a statistician.** Small-and-skewed is exactly the case where general-purpose tools fail and specific expertise helps.

Reporting a t-interval or a plain percentile CI in this situation without flagging the concern would overstate your confidence. See Hesterberg (2015), *"What Teachers Should Know About the Bootstrap"*, for the full accounting.

13.2.3 Worked example: Mercury levels in Risso's dolphins

Researchers measured mercury concentrations (in μg/g) in the muscle tissue of 19 Risso's dolphins from the Adriatic Sea. The sample mean was $\bar{x} = 4.4$ μg/g and the sample standard deviation was $s = 2.3$ μg/g. Construct a 95% confidence interval for the mean mercury level in the population of Risso's dolphins.

Check conditions.

1. *Independence.* The dolphins were sampled from different locations, so it is reasonable to treat the measurements as independent.
2. *Normality.* With $n = 19$, we need to check for strong skewness or outliers. Suppose a histogram of the data shows a roughly symmetric distribution with no extreme outliers. The condition is satisfied.

Compute the interval. With $df = 19 - 1 = 18$, the critical value for a 95% CI is $t^*_{18} = 2.101$.

$$SE = \frac{s}{\sqrt{n}} = \frac{2.3}{\sqrt{19}} = \frac{2.3}{4.359} = 0.528$$

$$\bar{x} \pm t^*_{18} \times SE = 4.4 \pm 2.101 \times 0.528 = 4.4 \pm 1.109 = (3.29, 5.51)$$

We are 95% confident that the mean mercury concentration in Risso's dolphins is between 3.29 and 5.51 μg/g.

A sample of 35 students at a college reported their weekly study hours. The sample mean is $\bar{x} = 15.2$ hours and $s = 6.8$ hours. Construct a 90% confidence interval for the mean weekly study hours of all students at that college. (Use $t^*_{34} = 1.691$ for a 90% CI.)

Show answer

First check conditions: the sample is a random sample (independence), and $n = 35 \geq 30$ (normality via CLT). Then $SE = 6.8/\sqrt{35} = 1.149$. The 90% CI is $15.2 \pm 1.691 \times 1.149 = 15.2 \pm 1.943 = (13.26, 17.14)$. We are 90% confident the mean weekly study hours for all students at that college is between 13.26 and 17.14 hours.

13.3 Two-sample t confidence interval

13.3.1 The setup

When comparing the means of two independent groups, we estimate the difference $\mu_1 - \mu_2$ using $\bar{x}_1 - \bar{x}_2$.

Two-sample t confidence interval for $\mu_1 - \mu_2$.

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Degrees of freedom: The exact formula for df is complex and typically computed by software. A conservative (safe) approximation is:

$$df = \min(n_1 - 1,\ n_2 - 1)$$

Conditions:

1. **Independence (extended).** Observations are independent within each group *and* between groups (e.g., from two independent random samples or a randomized experiment).
2. **Normality / sample size.** Each group's data should be approximately normal or have $n \geq 30$. Check each group separately for strong skewness or extreme outliers.

13.3.2 Worked example: Embryonic stem cells and heart function

Researchers studied whether embryonic stem cell (ESC) treatment improves heart function in sheep after a heart attack. Nine sheep received the ESC treatment and nine were in the control group. The change in pumping capacity was measured:

Table 13.2: Sample statistics for the stem cell study.

Group	n	$\bar{x}$	s
ESC	9	3.50	5.17
Control	9	-4.33	2.76

Construct a 95% confidence interval for the difference in mean change in pumping capacity, $\mu_{ESC} - \mu_{Control}$.

Check conditions.

1. *Independence.* The sheep were randomly assigned to groups, so independence within and between groups is satisfied.
2. *Normality.* With $n = 9$ per group (small samples), we check histograms. Suppose they show no extreme outliers in either group. The condition is reasonably met.

Compute the interval.

$$SE = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = \sqrt{\frac{26.73}{9} + \frac{7.62}{9}} = \sqrt{2.970 + 0.847} = \sqrt{3.817} = 1.954$$

Using $df = \min(9-1, 9-1) = 8$, the critical value for a 95% CI is $t^*_8 = 2.306$.

$$\text{Point estimate} = 3.50 - (-4.33) = 7.83$$

$$7.83 \pm 2.306 \times 1.954 = 7.83 \pm 4.506 = (3.32, 12.34)$$

We are 95% confident that the ESC treatment increases heart pumping capacity by between 3.32% and 12.34% compared to the control. Because the interval is entirely above zero, we have evidence that the ESC treatment improves heart function. Since this was a randomized experiment, we can attribute the improvement to the treatment.

Researchers collected birth weight data from a random sample of 1,000 US births. Nonsmoker mothers ($n = 867$) had babies with $\bar{x}_n = 7.27$ lbs ($s_n = 1.23$), and smoker mothers ($n = 114$) had babies with $\bar{x}_s = 6.68$ lbs ($s_s = 1.60$). Is there evidence that babies born to nonsmoker mothers weigh more on average? Compute a 95% confidence interval for the difference $\mu_n - \mu_s$.

Show answer

The point estimate is $7.27 - 6.68 = 0.59$ lbs. The SE is $\sqrt{1.23^2/867 + 1.60^2/114} = \sqrt{0.001745 + 0.02246} = \sqrt{0.02420} = 0.156$. Using $df = \min(866, 113) = 113$ and $t^*_{113} \approx 1.981$, the CI is $0.59 \pm 1.981 \times 0.156 = 0.59 \pm 0.309 = (0.28, 0.90)$. We are 95% confident that babies born to nonsmoker mothers weigh between 0.28 and 0.90 lbs more on average than babies born to smoker mothers.

13.4 Paired data confidence interval

13.4.1 What makes data "paired"?

Sometimes two measurements are made on the same individual (e.g., before and after a treatment) or on naturally matched pairs (e.g., the same textbook priced at two stores). In these cases, the two groups are **not independent** – they are **paired**.

Paired data. Two sets of observations are **paired** if each observation in one set has a special correspondence or connection with exactly one observation in the other set. Common examples include:

- Before/after measurements on the same individual
- Two treatments applied to the same subject (e.g., tires on the same car)
- Matched pairs in an experiment (e.g., twins assigned to different treatments)
- The same item measured under two conditions (e.g., textbook prices at two stores)

Table 13.3: Examples of paired designs.

Observational unit	Comparison groups	Measurement	Value of interest
Car	Smooth Turn vs. Quick Spin	Tire tread after 1,000 miles	Difference in tread
Textbook	UCLA Bookstore vs. Amazon	Price of new textbook	Difference in price

Observational unit	Comparison groups	Measurement	Value of interest
Individual	Pre-course vs. Post-course	Exam score	Difference in score

13.4.2 The key idea: analyze the differences

The key to working with paired data is simple: **compute the difference for each pair, then analyze those differences as a single sample.** This converts a two-sample problem into a one-sample problem.

Paired t confidence interval for μ_{diff}.

For paired data, compute the difference $d_i = x_{1,i} - x_{2,i}$ for each pair. Then the confidence interval for the mean difference μ_{diff} is:

$$\bar{x}_{diff} \pm t^*_{df} \times \frac{s_{diff}}{\sqrt{n_{diff}}}$$

where $\bar{x}_{diff}$ is the sample mean of the differences, s_{diff} is the sample standard deviation of the differences, n_{diff} is the number of pairs, and $df = n_{diff} - 1$.

Conditions:

1. **Independence of pairs.** The pairs themselves are independent (e.g., different cars, different textbooks).
2. **Normality of differences.** The distribution of the *differences* should be approximately normal, or the number of pairs should be large ($n_{diff} \geq 30$). Check for extreme outliers.

13.4.3 Worked example: Textbook prices

Researchers sampled 68 UCLA courses and compared textbook prices at the UCLA Bookstore to prices on Amazon. Sample statistics for the price differences (UCLA − Amazon) are:

Table 13.4: Sample statistics for the 68 price differences.

n_{diff}	$\bar{x}_{diff}$	s_{diff}
68	\$3.58	\$13.42

Construct a 95% confidence interval for the mean price difference.

Check conditions.

1. *Independence of pairs.* The textbooks were randomly sampled, so the pairs are independent.
2. *Normality of differences.* With $n_{diff} = 68$, the CLT ensures the sampling distribution of $\bar{x}_{diff}$ is approximately normal. Though the distribution of individual differences is right-skewed, the large sample size and absence of extreme outliers make the t-methods reasonable.

Compute the interval.

$$SE = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{13.42}{\sqrt{68}} = \frac{13.42}{8.246} = 1.628$$

With $df = 68 - 1 = 67$, the critical value for a 95% CI is $t^*_{67} \approx 2.00$.

$$3.58 \pm 2.00 \times 1.628 = 3.58 \pm 3.256 = (0.32, 6.84)$$

We are 95% confident that UCLA Bookstore prices are, on average, between \$0.32 and \$6.84 higher than Amazon prices for UCLA course books.

Try this in StatLens →

A fitness program enrolled 25 participants and measured their resting heart rates before and after 8 weeks of training. The mean difference (before − after) was $\bar{x}_{diff} = 4.2$ beats per minute with $s_{diff} = 6.1$ bpm. Construct a 90% confidence interval for the mean decrease in resting heart rate. (Use $t^*_{24} = 1.711$.)

Show answer

First check conditions: pairs (individuals) were randomly selected (independence), and with $n = 25$ and no extreme outliers reported, the normality condition is reasonable. Then $SE = 6.1/\sqrt{25} = 6.1/5 = 1.22$. The 90% CI is $4.2 \pm 1.711 \times 1.22 = 4.2 \pm 2.09 = (2.11, 6.29)$. We are 90% confident the training program reduces resting heart rate by between 2.11 and 6.29 bpm on average.

13.4.4 Why not use a two-sample t interval for paired data?

Never use a two-sample procedure for paired data. A two-sample t interval treats the two groups as independent, which they are not. Ignoring the pairing throws away valuable information about the within-pair relationship and typically produces a wider (less precise) confidence interval. Always compute the differences first and then use the one-sample (paired) t procedure.

13.5 Connecting t-based and bootstrap confidence intervals

The t-based confidence intervals in this chapter and the bootstrap confidence intervals from the bootstrap confidence intervals chapter are answering the same question: "What is a plausible range for the population parameter?" When conditions are met, they give similar results.

When do they agree?

- Large samples with no extreme outliers: the t-interval and bootstrap interval will be nearly identical.
- Moderate samples from approximately normal populations: close agreement.

When might they differ?

- Small samples from skewed populations: the bootstrap can capture asymmetry in the sampling distribution, while the t-interval is always symmetric around $\bar{x}$.
- Presence of outliers: the bootstrap is somewhat more robust because it does not rely on s being a good estimate of σ.

Which should you use?

- If conditions are met, the t-interval is simpler to compute and has well-understood properties.
- If conditions are questionable (small sample, skewness, outliers), prefer the bootstrap.
- In practice, reporting both and noting any discrepancies is informative.

13.6 Chapter review

13.6.1 Summary

This chapter developed formula-based confidence intervals for means using the t-distribution.

- The **t-distribution** accounts for the extra uncertainty from estimating σ with s. It has heavier tails than the normal distribution, especially for small degrees of freedom.
- **One-sample t CI**: $\bar{x} \pm t^*_{n-1} \times \frac{s}{\sqrt{n}}$. Requires independent observations and approximate normality (or large n).
- **Two-sample t CI**: $(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. Requires independence within and between groups, and approximate normality or large samples in each group.
- **Paired t CI**: Compute differences for each pair, then apply the one-sample t CI to the differences: $\bar{x}_{diff} \pm t^*_{n_{diff}-1} \times \frac{s_{diff}}{\sqrt{n_{diff}}}$.
- Always check conditions before using t-methods: independence, normality/sample size, and absence of extreme outliers.
- When conditions are met, t-based and bootstrap confidence intervals give similar results. When conditions are questionable, the bootstrap may be more reliable.

13.6.2 Key terms

t-distribution, degrees of freedom (df), critical value (t^*), one-sample t confidence interval, two-sample t confidence interval, paired data, paired t confidence interval, standard error of the mean, independence condition, normality condition.

13.7 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Statistics vs. parameters: one mean.** Each of the following scenarios were set up to assess an average value. For each one, identify, in words: the statistic and the parameter.

a. A sample of 25 New Yorkers were asked how much sleep they get per night.

b. Researchers at two different universities in California collected information on undergraduates' heights.

2. **Statistics vs. parameters: one mean.** Each of the following scenarios were set up to assess an average value. For each one, identify, in words: the statistic and the parameter.

a. Georgianna samples 20 children from a particular city and measures how many years they have each been playing piano.

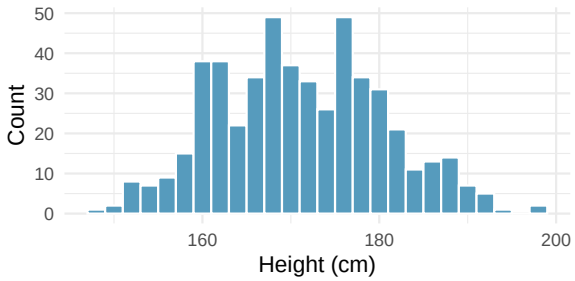
b. Traffic police officers (who are regularly exposed to lead from automobile exhaust) had their lead levels measured in their blood.

3. **Heights of adults.** Researchers studying anthropometry collected body measurements, as well as age, weight, height and gender, for 507 physically active adults. Summary statistics for the distribution of heights (measured in centimeters, cm), along with a histogram, are provided below. (Heinz et al. 2003)

a. What are the point estimates for the average and median heights of active adults?

b. What are the point estimates for the standard deviation and IQR of heights of active adults?

Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
Q3	177.8
Max	198.1
SD	9.4
IQR	14.0



c. Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

d. The researchers take another random sample of physically active adults. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

e. The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

4. **Heights of adults, standard error.** Heights of 507 physically active adults have a mean of 171 cm and a standard deviation of 9.4 cm. Provide an estimate for the standard error of the mean for samples of following sizes. (Heinz et al. 2003)

a. n = 10

b. n = 50

c. n = 100

d. n = 1000

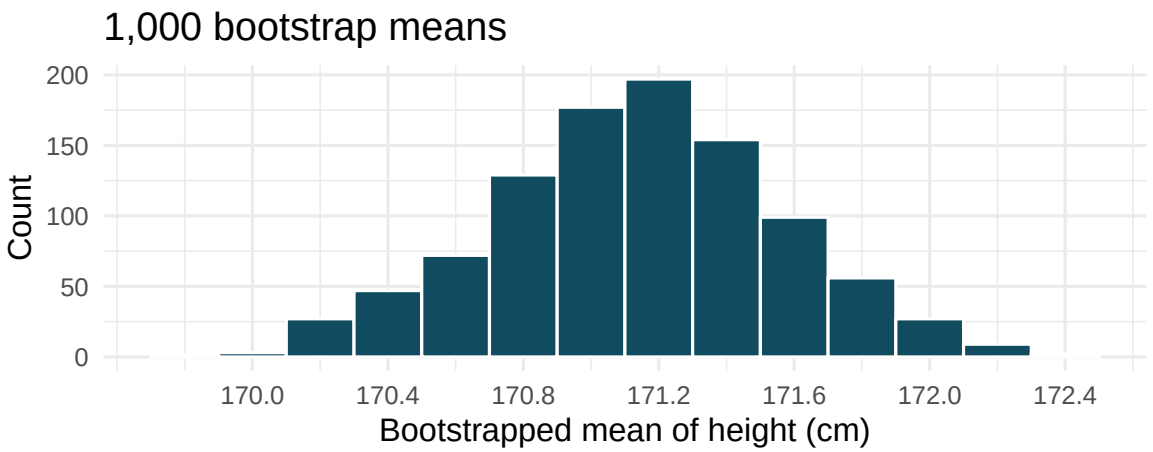
e. The standard error of the mean is a number which describes what?

5. **Heights of adults vs. kindergartners.** Heights of 507 physically active adults have a mean of 171 cm and a standard deviation of 9.4 cm. (Heinz et al. 2003)

a. Would you expect the standard deviation of the heights of a few hundred kindergartners to be higher or lower than 9.4 cm? Explain your reasoning.

b. Suppose many samples of size 100 adults is taken and, separately, many samples of size 100 kindergarteners are taken. For each of the many samples, the average height is computed. Which set of sample averages would have a larger standard error of the mean, the adult sample averages or the kindergartner sample averages?

6. **Heights of adults, bootstrap interval.** Researchers studying anthropometry collected body measurements, as well as age, weight, height and gender, for 507 physically active adults. The histogram below shows the sample distribution of bootstrapped means from 1,000 different bootstrap samples. (Heinz et al. 2003)



a. Given the bootstrap sampling distribution for the sample mean, find an approximate value for the standard error of the mean.

b. By looking at the bootstrap sampling distribution (1,000 bootstrap samples were taken), find an approximate 90% bootstrap percentile confidence interval for the true average adult height in the population from which the data were randomly sampled. Provide the interval as well as a one-sentence interpretation of the interval.

c. By looking at the bootstrap sampling distribution (1,000 bootstrap samples were taken), find an approximate 90% bootstrap SE confidence interval for the true average adult height in the population from which the data were randomly sampled. Provide the interval as well as a one-sentence interpretation of the interval.

7. **Identify the critical t.** A random sample is selected from an approximately normal population with unknown standard deviation. Find the degrees of freedom and the critical t-value ($t^\star$) for the given sample size and confidence level.

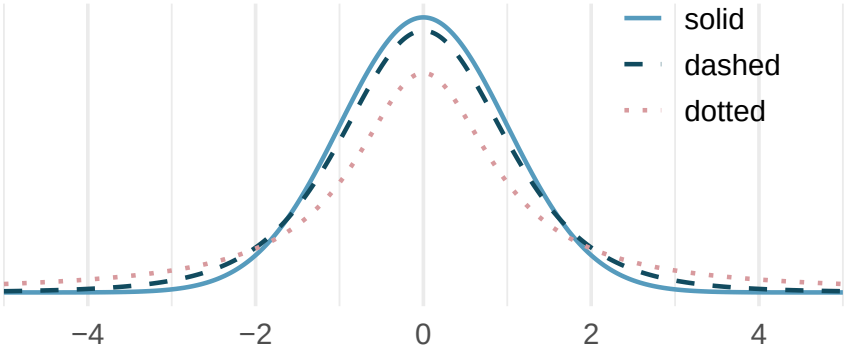
a. $n = 6$, CL = 90%

b. $n = 21$, CL = 98%

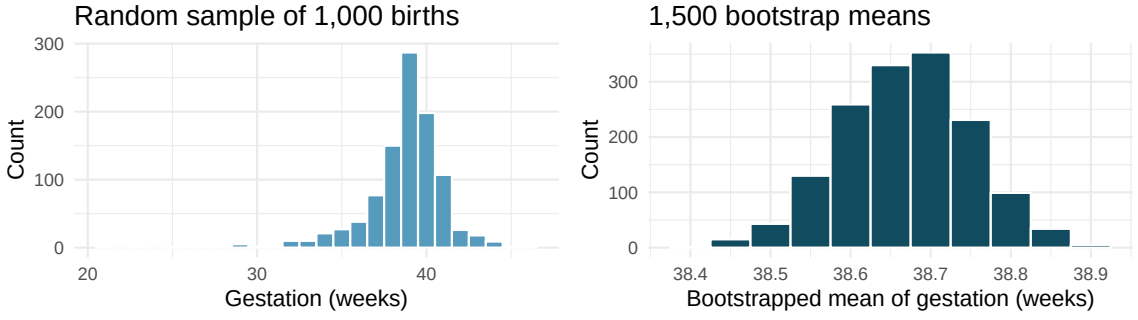
c. $n = 29$, CL = 95%

d. $n = 12$, CL = 99%

8. **t-distribution.** The figure below shows three unimodal and symmetric curves: the standard normal (z) distribution, the t-distribution with 5 degrees of freedom, and the t-distribution with 1 degree of freedom. Determine which is which, and explain your reasoning.



9. **Length of gestation, confidence interval.** Every year, the United States Department of Health and Human Services releases to the public a large dataset containing information on births recorded in the country. This dataset has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. In this exercise we work with a random sample of 1,000 cases from the dataset released in 2014. The length of pregnancy, measured in weeks, is commonly referred to as gestation. The histograms below show the distribution of lengths of gestation from the random sample of 1,000 births (on the left) and the distribution of bootstrapped means of gestation from 1,500 different bootstrap samples (on the right).



a. Given the bootstrap sampling distribution for the sample mean, find an approximate value for the standard error of the mean.

b. By looking at the bootstrap sampling distribution (1,500 bootstrap samples were taken), find an approximate 99% bootstrap percentile confidence interval for the true average gestation length in the population from which the data were randomly sampled. Provide the interval as well as a one-sentence interpretation of the interval.

c. By looking at the bootstrap sampling distribution (1,500 bootstrap samples were taken), find an approximate 99% bootstrap SE confidence interval for the true average gestation length in the population from which the data were randomly sampled. Provide the interval as well as a one-sentence interpretation of the interval.

10. **Interpreting confidence intervals for population mean.** For each of the following statements, indicate if they are a true or false interpretation of the confidence interval. If false, provide a reason or correction to the misinterpretation. You collect a large sample and calculate a 95% confidence interval for the average number of cans of sodas consumed annually per adult in the US to be (440 cans, 520 cans), i.e., on average, adults in the US consume just under two cans of soda per day.

a. 95% of adults in the US consume between 440 and 520 cans of soda per year.

b. There is a 95% probability that the true population average per adult yearly soda consumption is between 440 and 520 cans.

c. The true population average per adult yearly soda consumption is between 440 and 520 cans, with 95% confidence.

d. The average soda consumption of the people who were sampled is between 440 and 520 cans of soda per year, with 95% confidence.

11. **Working backwards, I.** A 95% confidence interval for a population mean, μ, is given as (18.985, 21.015). The population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 36 observations. Assuming that all conditions necessary for inference are satisfied, and using the t-distribution, calculate the sample mean, the margin of error, and the sample standard deviation.

12. **Working backwards, II.** A 90% confidence interval for a population mean is (65, 77). The population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 25 observations. Assuming that all conditions necessary for inference are satisfied, and using the t-distribution, calculate the sample mean, the margin of error, and the sample standard deviation.

13. $t^\star$ **for the correct confidence level.** As you've seen, the tails of a $t-$distribution are longer than the standard normal which results in $t^\star_{df}$ being larger than $z^\star$ for any given confidence level. When finding a CI for a population mean, explain how mistakenly using $z^\star$ (instead of the correct t^*_{df}) would affect the confidence level.

14. **Possible bootstrap samples.** Consider a simple random sample of the following observations: 47, 4, 92, 47, 12, 8. Which of the following could be a possible bootstrap samples from the observed data above? If the set of values could not be a bootstrap sample, indicate why not.

a. 47, 47, 47, 47, 47, 47

b. 92, 4, 13, 8, 47, 4

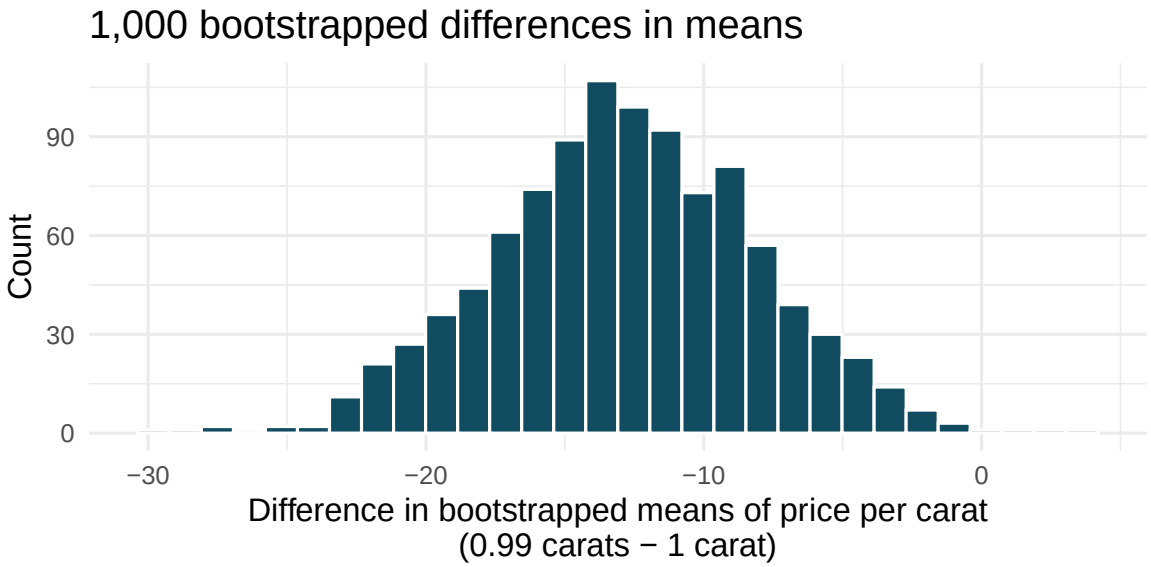
c. 92, 47, 12

d. 8, 47, 12, 12, 8, 4, 92

e. 12, 4, 8, 8, 92, 12

15. **Fill in the blanks.** We use a ___ to evaluate if data provide convincing evidence of a difference between two population means and we use a ___ to estimate this difference.

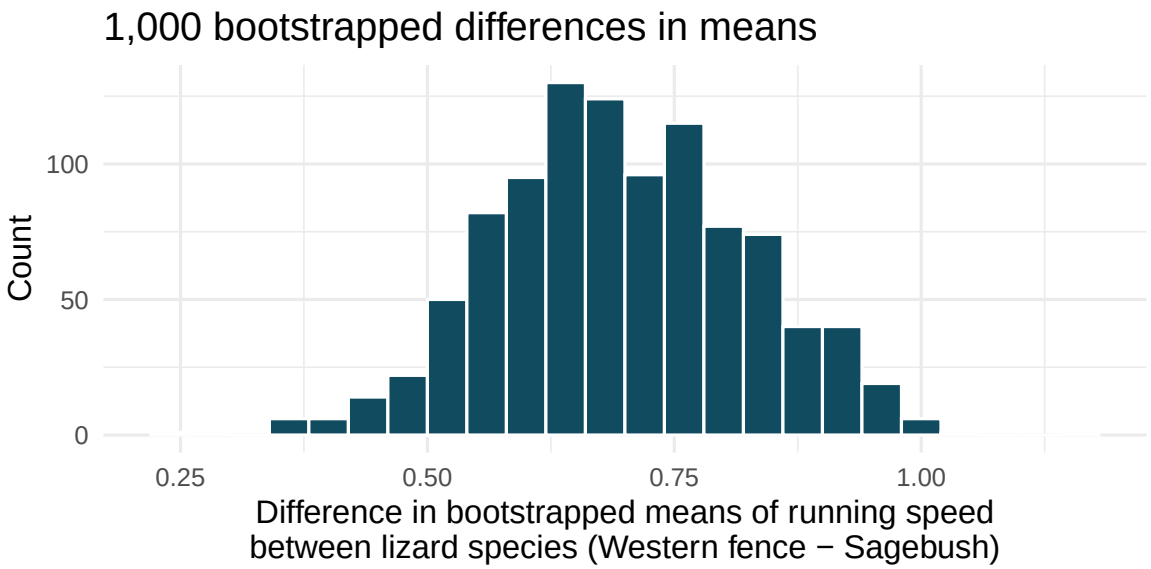
16. **Diamonds, bootstrap interval.** We have data on two random samples of diamonds: 23 0.99 carat diamonds and 23 1 carat diamonds. Provided below is a histogram of bootstrap differences in means of price per carat of diamonds that weigh 0.99 carats and diamonds that weigh 1 carat. (Wickham 2016)



a. Using the bootstrap distribution, create a (rough) 95% bootstrap percentile confidence interval for the true population difference in prices per carat of diamonds that weigh 0.99 carats and 1 carat.

b. Using the bootstrap distribution, create a (rough) 95% bootstrap SE confidence interval for the true population difference in prices per carat of diamonds that weigh 0.99 carats and 1 carat. Note that the standard error of the bootstrap distribution is 4.64.

17. **Lizards running, bootstrap interval.** We have data on top speeds (in m/sec) measured on a laboratory race track for two species of lizards: Western fence lizard (Sceloporus occidentalis) and Sagebrush lizard (Sceloporus graciosus). The bootstrap distribution below describes the variability of difference in means captured from 1,000 bootstrap samples of the lizard data. (Adolph 1987)



a. Using the bootstrap distribution, create a (rough) 90% percentile bootrap confidence interval for the true population difference in average speed of the Western fence lizard as compared with Sagebrush lizard.

b. Using the bootstrap distribution, create a (rough) 90% bootstrap SE confidence interval for the true population difference in average speed of the Western fence lizard as compared with Sagebrush lizard.

18. **Weight loss.** You are reading an article in which the researchers have created a 95% confidence interval for the difference in average weight loss for two diets. They are 95% confident that the true difference in average weight loss over 6 months for the two diets is somewhere between (1 lb, 25 lbs). The authors claim that, "therefore diet A ($\bar{x}_A$ = 20 lbs average loss) results in a much larger average weight loss as compared to diet B ($\bar{x}_B$ = 7 lbs average loss)." Comment on the authors' claim.

19. **Diamonds, mathematical interval.** We have data on two random samples of diamonds: one with diamonds that weigh 0.99 carats and one with diamonds that weigh 1 carat. Each sample has 23 diamonds. Sample statistics for the price per carat of diamonds in each sample are provided below. Assuming that the conditions for conducting inference using a mathematical model are satisfied, construct a 95% confidence interval for the true population difference in prices per carat of diamonds that weigh 0.99 carats and 1 carat. (Wickham 2016)

	Mean	SD	n
0.99 carats	\$44.51	\$13.32	23
1 carat	\$57.20	\$18.19	23

20. **True / False: comparing means.** Determine if the following statements are true or false, and explain your reasoning for statements you identify as false.

a. As the degrees of freedom increases, the t-distribution approaches normality.

b. If a 95% confidence interval for the difference between two population means contains 0, a 99% confidence interval calculated based on the same two samples will also contain 0.

c. If a 95% confidence interval for the difference between two population means contains 0, a 90% confidence interval calculated based on the same two samples will also contain 0.

21. **Difference of means.** We collect two random samples from two different populations In each part below, consider the sample means $\bar{x}_1$ and $\bar{x}_2$ that we might observe from these two samples.

	Mean	Standard deviation	Sample size
Population 1	15	20	50
Population 2	20	10	30

a. What is the associated mean and standard deviation of $\bar{x}_1$?

b. What is the associated mean and standard deviation of $\bar{x}_2$?

c. Calculate and interpret the mean and standard deviation associated with the difference in sample means for the two groups, $\bar{x}_2 - \bar{x}_1$.

d. How are the standard deviations from parts (a), (b), and (c) related?

22. **Air quality.** Air quality measurements were collected in a random sample of 25 country capitals in 2013, and then again in the same cities in 2014. We would like to use these data to compare average air quality between the two years. Should we use a paired or non-paired test? Explain your reasoning.

23. **True / False: paired.** Determine if the following statements are true or false. If false, explain.

a. In a paired analysis we first take the difference of each pair of observations, and then we do inference on these differences.

b. Two datasets of different sizes cannot be analyzed as paired data.

c. Consider two sets of data that are paired with each other. Each observation in one dataset has a natural correspondence with exactly one observation from the other dataset.

d. Consider two sets of data that are paired with each other. Each observation in one dataset is subtracted from the average of the other dataset's observations.

24. **Paired or not? I.** In each of the following scenarios, determine if the data are paired.

a. Compare pre- (beginning of semester) and post-test (end of semester) scores of students.

b. Assess gender-related salary gap by comparing salaries of randomly sampled men and women.

c. Compare artery thicknesses at the beginning of a study and after 2 years of taking Vitamin E for the same group of patients.

d. Assess effectiveness of a diet regimen by comparing the before and after weights of subjects.

25. **Paired or not? II.** In each of the following scenarios, determine if the data are paired.

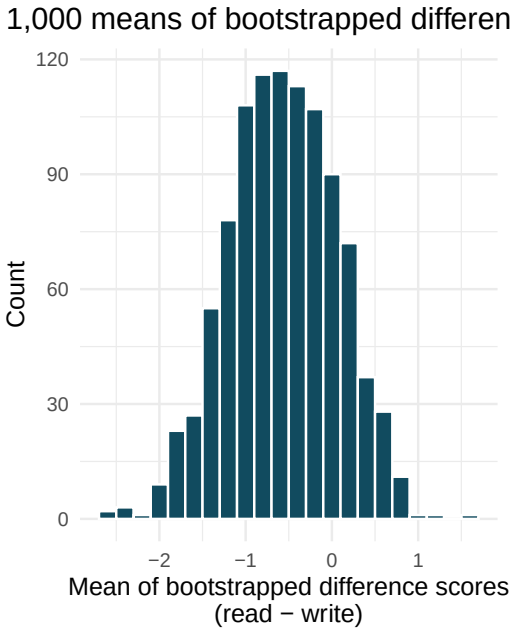
a. We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days, and record Intel's and Southwest's stock on those same days.

b. We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items.

c. A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.

26. **Sample size and pairing.** Determine if the following statement is true or false, and if false, explain your reasoning: If comparing means of two groups with equal sample sizes, always use a paired test.

27. **High School and Beyond, bootstrap interval.** We considered the differences between the reading and writing scores of a random sample of 200 students who took the High School and Beyond Survey. The mean and standard deviation of the differences are $\bar{x}_{read-write} = -0.545$ and $s_{read-write}$ = 8.887 points. The bootstrap distribution below was produced by bootstrapping from the sample of differences in reading and writing scores 1,000 times.



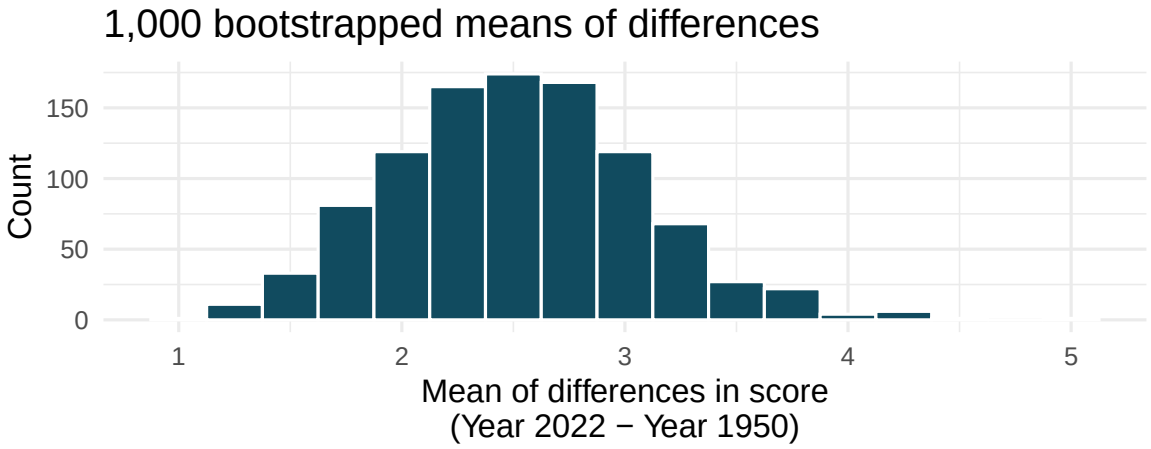
a. Find an approximate 95% bootstrap percentile confidence interval for the true average difference in scores (read - write).

b. Find an approximate 95% bootstrap SE confidence interval for the true average difference in scores (read - write).

c. Interpret both confidence intervals using words like "population" and "score".

d. From the confidence intervals calculated above, does it appear that there is a discernible difference in reading and writing scores, on average?

28. **Global warming, bootstrap interval.** We considered the change in the 90th percentile high temperature in 1950 versus 2022 at 26 sampled locations from the NOAA database. (NOAA 2023) The mean and standard deviation of the reported differences are 2.53°F and 2.95°F.



a. Calculate a 90% bootstrap percentile confidence interval for the average difference of 90th percentile high temperature between 1950 and 2022.

b. Calculate a 90% bootstrap SE confidence interval for the average difference of 90th percentile high temperature between 1950 and 2022.

c. Interpret both intervals in context.

d. Do the confidence intervals provide convincing evidence that there were hotter high temperatures in 2022 than in 1950 at NOAA stations? Explain your reasoning.

29. **High school and beyond, mathematical interval.** We considered the differences between the reading and writing scores of a random sample of 200 students who took the High School and Beyond Survey. The mean and standard deviation of the differences are $\bar{x}_{read-write} = -0.545$ and $s_{read-write}$ = 8.887 points.

a. Calculate a 95% confidence interval for the average difference between the reading and writing scores of all students.

b. Interpret this interval in context.

c. Does the confidence interval provide convincing evidence that there is a real difference in the average scores? Explain.

30. **Global warming, mathematical interval.** We considered the change in the 90th percentile high temperature in 1950 versus 2022 at 26 sampled locations from the NOAA database. (NOAA 2023) The mean and standard deviation of the reported differences are 2.53°F and 2.95°F.

a. Calculate a 90% confidence interval for the average difference of 90th percentile high temperature between 1950 and 2022. We've already checked the conditions for you.

b. Interpret the interval in context.

c. Does the confidence interval provide convincing evidence that there were hotter high temperatures in 2022 than in 1950 at NOAA stations? Explain your reasoning.

31. **Forest management.** Forest rangers wanted to better understand the rate of growth for younger trees in the park. They took measurements of a random sample of 50 young trees in 2009 and again measured those same trees in 2019. The data below summarize their measurements, where the heights are in feet.

Year	Mean	SD	n
2009	12.0	3.5	50
2019	24.5	9.5	50
Difference	12.5	7.2	50

Construct a 99% confidence interval for the average growth of (what had been) younger trees in the park over 2009-2019.

Datasets: `bdims` (openintro) | `births14` (openintro)

StatLens Exercises

These exercises focus on the parts of inference a calculator *can't* do for you: **framing the question**, **choosing the right procedure**, **checking conditions**, and **interpreting** what a tool reports. Where a link is given, let StatLens do the arithmetic and spend your effort on the reasoning.

1. **Frame the question.** A nutrition researcher wants to estimate the *mean* daily caffeine intake (in mg) among undergraduates at a large university. She collects a random sample of $n = 48$ students and records each one's daily intake.

   a. What is the parameter of interest? Describe it in words and give its symbol.
   b. Should the resulting confidence interval be **one-sided** or **two-sided**, and why?
   c. Why is the **sample standard deviation** s used in the t-procedure here, and not the **population standard deviation** σ?

2. **Choose the procedure.** For each scenario, name the procedure you would use — **one-sample** t (test or interval), **two-sample** t (independent groups), **paired** t (matched pairs), or **not a means problem**. You do *not* need to carry anything out; the skill here is matching the question to the method.

   a. Estimate the mean fuel economy (mpg) of a new car model from a sample of 25 test drives.
   b. Compare the mean recovery time of patients on Drug A vs. Drug B, where each patient was randomly assigned to *one* drug.
   c. Measure each runner's 5K time before and after a 6-week training program; ask whether the program lowers times on average.
   d. Estimate the **proportion** of voters in a district who support a ballot measure.
   e. Compare the *median* household income in two counties.

3. **Check conditions before you trust the interval.** A small clinical study reports the resting heart rates of $n = 12$ adults after an experimental intervention. Open the data in the One-Sample t tool (pick any small numeric dataset, or paste in a small sample of your own) and inspect the distribution.

   a. What two conditions does the one-sample t-interval require? State each in plain language.
   b. With only $n = 12$, how would you decide whether the normality condition is reasonable? What feature of the dotplot or histogram would make you uneasy?
   c. If the histogram shows a clear right skew with one extreme outlier, is the t-interval still trustworthy? What would you do instead?

4. **Run it and interpret it (two means).** A study compared mean reaction times (ms) between a caffeinated group and a placebo group. Open a two-group dataset of your choice in the Two-Sample t tool.

   a. Identify the explanatory and response variables, and define the parameter $\mu_1 - \mu_2$ in words.

   b. Run the tool. Report the 95% confidence interval for $\mu_1 - \mu_2$. Does the interval contain 0?

   c. Write a one-sentence conclusion **in context**, and say whether your confidence interval and a two-sided hypothesis test at $\alpha = 0.05$ would lead to the same decision.

   d. Which statement is the correct interpretation of the interval in (b)?

      i. There is a 95% probability that $\mu_1 - \mu_2$ lies in this interval.
      ii. If the experiment were repeated many times, about 95% of the intervals built this way would contain the true difference $\mu_1 - \mu_2$.
      iii. 95% of caffeinated participants fall within this interval.

5. **Paired vs. independent — choose carefully, then compare.** Twelve runners each ran a 5K race **before** and **after** a 6-week training program; you also have a separate sample of 12 *different* runners who did not train, run at the same point in the season.

   a. If you want to know whether the *program* lowered 5K times, why is the paired-sample analysis (on the within-runner differences) almost always more informative than treating the before-times and after-times as two independent groups?
   b. Open the Paired t tool with any paired dataset. Record the 95% CI for the mean difference. Then re-analyze the *same* numbers in the Two-Sample t tool as if they were independent groups. What is different about the standard error and the width of the CI?
   c. Now consider the *untrained* sample. To estimate the program's effect against a no-training baseline, what would you compare, and which procedure would you use? Why is matching impossible here?
   d. A student says: "Paired is just a special case of two-sample." Why is this misleading, even though both end up using the t distribution?

6. **Compute-unit capstone — write conclusions across every procedure you have learned so far.** By this point you have worked with the one-proportion z, the two-proportion z, the one-sample t, the two-sample t, and the paired t. The last skill — taking a test result and *writing what it means in context* — is one that gets rusty quickly if you don't practice across procedures.

   Open Conclusion Practice — Compute scope. This link filters the scenario pool to just the five procedures from Part III (no chi-square, ANOVA, or regression — those come later). Each round gives you a randomized scenario with the test output already computed; your job is to assemble a **formal** conclusion (about the parameter, at the stated α) and a **practical** conclusion (what a non-statistician stakeholder should take away). The tool scores you.

   a. Play at least **5 rounds**. Aim for a passing score in each. Do you notice any pattern in the mistakes you make — do they cluster around a particular procedure, or a particular pitfall (misinterpreting p, misinterpreting the CI, over-claiming causation, etc.)?
   b. Which is harder for you: writing the **formal** conclusion, or writing the **practical** one? Why?
   c. Reflect: from the perspective of "what makes an inference class valuable," which mattered more — computing p-values (which StatLens did for you) or writing correct conclusions (which you had to do)? What does this suggest about how you should split your study time going forward?

Chapter 14

Hypothesis Tests for Means

In the randomization tests chapter, we developed the logic of hypothesis testing using simulation: shuffle the data under the null hypothesis, build a null distribution, and measure how extreme the observed statistic is. In the confidence intervals for means chapter, we introduced the t-distribution as a mathematical model for the sampling distribution of means. Now we bring these ideas together. In this chapter, we use the t-distribution to conduct hypothesis tests for a single mean, the difference between two independent means, and the mean of paired differences. Along the way, we will see that the t-test and simulation-based tests generally agree when conditions are met, and we will discuss what to do when they disagree.

14.1 Testing a single mean by simulation

In the randomization tests chapter we tested a *difference between two groups* by shuffling the group labels, and in the inference for proportions chapter we tested a *single proportion* by drawing from a modeled null population — the marbles bag. A single **mean** needs its own version of that idea. There are no group labels to shuffle, and the data are numbers rather than successes and failures. But the goal is the same: build a world in which the null hypothesis is exactly true, then see how often chance alone produces a sample mean as extreme as the one we observed.

Suppose we want to test $H_0 : \mu = \mu_0$. Our sample mean $\bar{x}$ almost certainly differs from μ_0 — the question is whether that difference is larger than sampling variability would produce if μ_0 really were the truth. We build the null world in two moves:

1. **Shift** the data so its mean is exactly μ_0. Add the same constant to every value (whatever it takes to move $\bar{x}$ onto μ_0). This forces the null hypothesis to be true while leaving the data's shape and spread untouched.
2. **Resample** with replacement from the shifted data, many times, recording each resample's mean. Those simulated means show how $\bar{x}$ would bounce around from sample to sample *if* $\mu = \mu_0$.

A bakery advertises that its sourdough loaves weigh **500 grams** on average. A consumer group suspects the loaves are underfilled, weighs a random sample of 15 loaves, and finds a sample mean of **492.8 grams** — about 7 grams light. Could a sample this light happen by chance when the true mean is 500 g?

We test $H_0 : \mu = 500$ against $H_A : \mu < 500$. To simulate the null, we shift the 15 weights up so their mean is exactly 500, then repeatedly resample 15 loaves (with replacement) and record the mean of each.

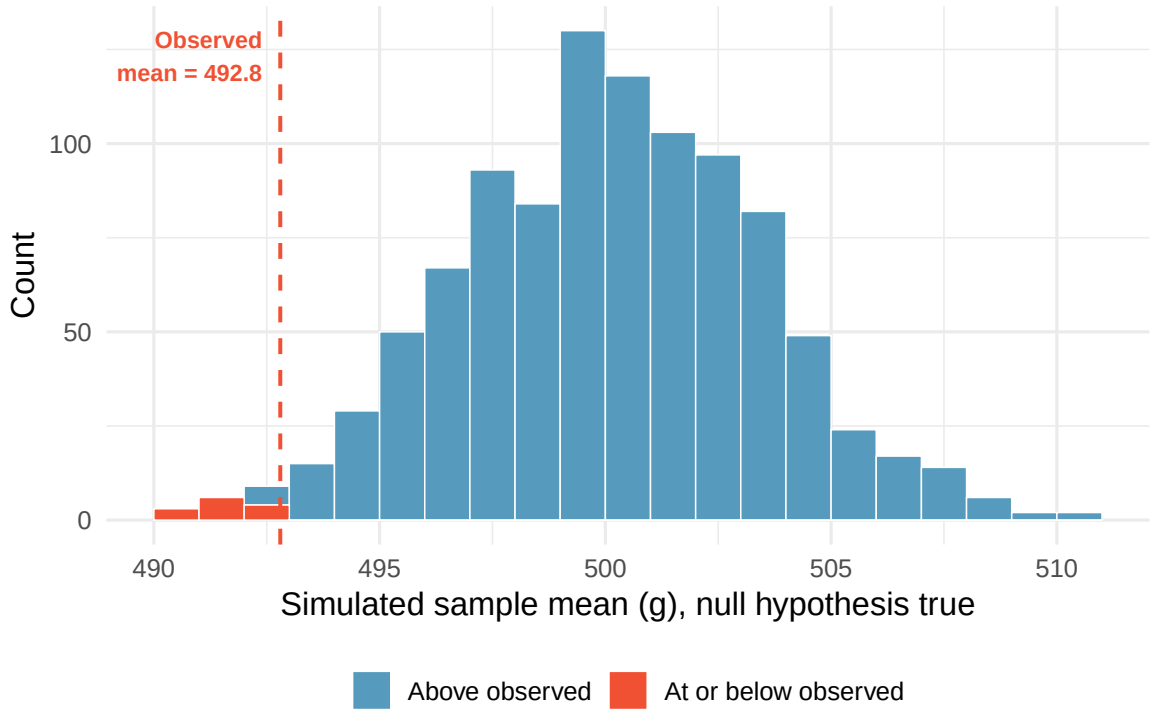


Figure 14.1: A histogram of 1,000 simulated sample means, each from a sample of 15 loaves drawn from data shifted to satisfy the null hypothesis (mean = 500 g). Simulated means at or below the observed 492.8 g are shaded in red.

The simulated means center on 500 g, as they must — we built the null world to make that true. The observed 492.8 g falls well into the left tail: a sample mean this low turns up only about 1% of the time when the true mean is really 500 g. That small p-value is evidence the loaves are underfilled.

Notice what this simulation assumed: only that the null is true and the sample is representative. No bell curve, no formula. The rest of this chapter develops the **t-test** — a mathematical shortcut that gives essentially the same answer when the data are roughly symmetric and the sample is not too small, and that comes with a formula you can compute by hand.

See it in action. Open the Randomization Test for a Single Mean to run this shift-and-resample test on your own data — enter a sample and a null value, generate the null distribution, and read the p-value off the tail.

14.2 The t-test framework

Every hypothesis test follows the same four-step structure — the same one we used for proportions — whether we use simulation or a mathematical model:

1. **State hypotheses.** Write H_0 (null hypothesis) and H_A (alternative hypothesis) in both words and symbols.
2. **Check conditions.** Verify that the mathematical model is appropriate.
3. **Compute the test statistic and p-value.** Calculate how far the observed data are from what H_0 predicts, and determine how unusual that distance is.
4. **Draw a conclusion.** Compare the p-value to the discernibility level α and state the conclusion in context.

For means, the test statistic is a T **score** – the standardized distance from the observed sample mean to the hypothesized value, measured in units of standard errors.

See it in action. Guided walkthroughs for each t-test: one-sample · two-sample · paired. Each gates you through framing → conditions → test → interpretation. Or open the corresponding analytical CI + test tool directly on your own data.

14.3 One-sample t-test

14.3.1 Setup and hypotheses

The one-sample t-test is used when we have a single sample of quantitative data and want to test a claim about the population mean μ.

The test statistic for a single mean is a T.

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean, μ_0 is the hypothesized value from H_0, s is the sample standard deviation, and n is the sample size.

When the null hypothesis is true and the conditions are met, T follows a t-distribution with $df = n - 1$.

Conditions:

- **Independent observations** (e.g., random sample).
- **Normality / sample size:** The population distribution should be approximately normal, or n should be large enough ($n \geq 30$). Check for strong skewness or extreme outliers.

14.3.2 Worked example: Cherry Blossom race times

The average finishing time for all runners in the Cherry Blossom 10-mile race in 2006 was $\mu = 93.29$ minutes. A random sample of 100 runners from a more recent year had a mean finishing time of $\bar{x} = 97.32$ minutes with $s = 16.98$ minutes. Is there evidence that the average finishing time has changed?

Step 1: State hypotheses.

- H_0: $\mu = 93.29$ (the average finishing time has not changed)
- H_A: $\mu \neq 93.29$ (the average finishing time has changed)

This is a two-sided test because we are looking for a change in either direction.

Step 2: Check conditions.

- *Independence.* The sample is a random sample of runners, so observations are independent.
- *Sample size.* With $n = 100 \geq 30$, the CLT guarantees the sampling distribution of $\bar{x}$ is approximately normal, even if the population distribution of finishing times is skewed.

Step 3: Compute the test statistic and p-value.

$$SE = \frac{s}{\sqrt{n}} = \frac{16.98}{\sqrt{100}} = 1.698$$

$$T = \frac{\bar{x} - \mu_0}{SE} = \frac{97.32 - 93.29}{1.698} = \frac{4.03}{1.698} = 2.37$$

With $df = 100-1 = 99$, we find the area to the right of $T = 2.37$ on the t-distribution: approximately 0.0099. Since this is a two-sided test, the p-value is $2 \times 0.0099 = 0.0198$.

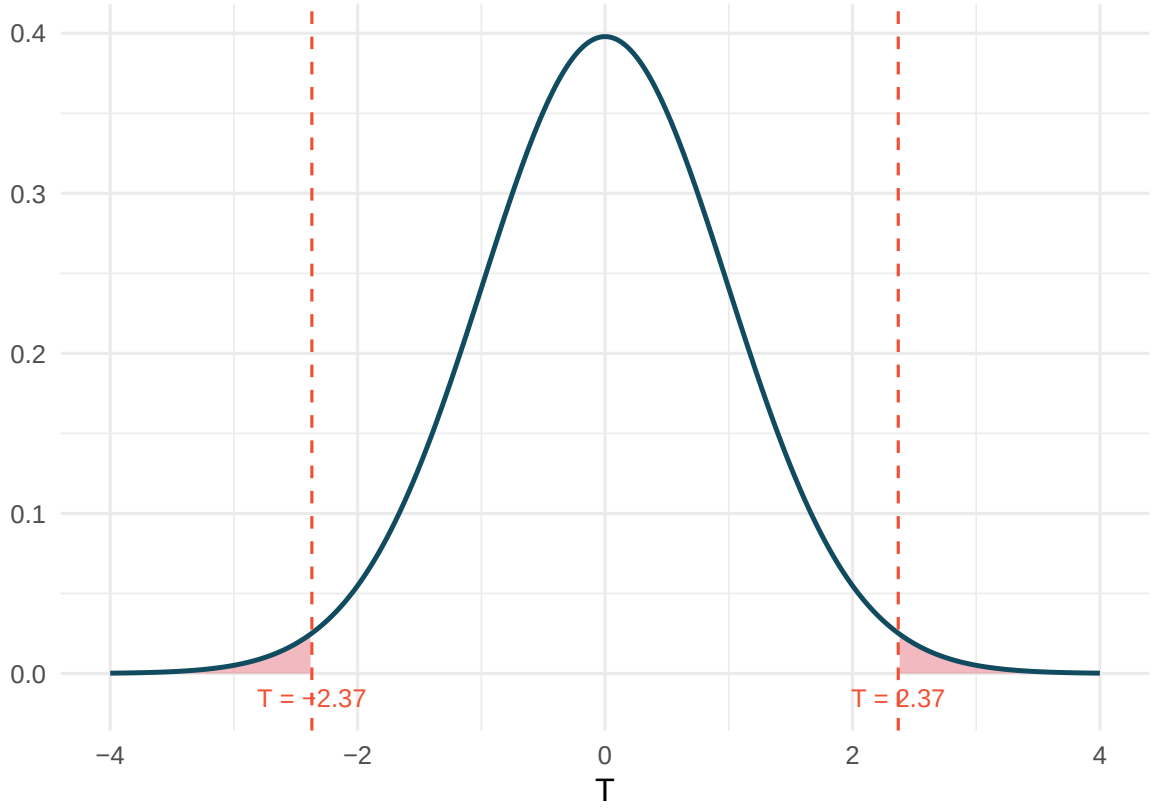


Figure 14.2: A t-distribution with df = 99 showing the two-sided p-value. The shaded areas in both tails represent the probability of observing a test statistic at least as extreme as 2.37 in either direction. ↗ Try this live — drag the cutlines to see the two-tail p-value change with t.

Step 4: Conclusion. The p-value of 0.0198 is less than $\alpha = 0.05$, so we reject H_0. The data provide statistically discernible evidence that the average finishing time has changed from the 2006 average of 93.29 minutes. Based on the sample, runners appear to be finishing more slowly on average.

Try this in StatLens →

A nutrition study claims that the average American adult consumes 2,000 mg of sodium per day. A random sample of 45 adults shows $\bar{x} = 3,120$ mg and $s = 1,100$ mg. Test whether the data provide evidence that the true average sodium intake differs from 2,000 mg. Use $\alpha = 0.05$.

Show answer

Hypotheses: $H_0 : \mu = 2000$, $H_A : \mu \neq 2000$. Conditions: random sample (independence), $n = 45 \geq 30$ (normality via CLT). Test statistic: $SE = 1100/\sqrt{45} = 163.9$, $T = (3120 - 2000)/163.9 = 6.83$. With $df = 44$, the two-tailed p-value is essentially 0 (far less than 0.05). We reject H_0. There is very strong evidence that the average sodium intake is different from (and appears much higher than) 2,000 mg per day.

14.4 Two-sample t-test

14.4.1 Setup and hypotheses

The two-sample t-test compares the means of two independent groups. The parameter of interest is $\mu_1 - \mu_2$, and the null hypothesis typically claims no difference.

The test statistic for comparing two independent means is a T.

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

When the null hypothesis is true and conditions are met, T follows a t-distribution with $df = \min(n_1 - 1, n_2 - 1)$ (conservative approximation; software computes the exact df).

Conditions:

- **Independence (extended).** Observations are independent within and between the two groups.
- **Normality / sample size.** Each group should be approximately normal or have a large enough sample size. Check each group separately for strong skewness or extreme outliers.

14.4.2 Worked example: Birth weights and smoking

Is there evidence that newborns of mothers who smoke during pregnancy have a different average birth weight than newborns of non-smoking mothers? Data from a random sample of 1,000 US births in 2014:

Table 14.1: Sample statistics for birth weights by smoking status.

Group	n	Mean (lbs)	SD (lbs)
Nonsmoker	867	7.27	1.23
Smoker	114	6.68	1.60

Step 1: State hypotheses.

- H_0: $\mu_n - \mu_s = 0$ (no difference in average birth weight)
- H_A: $\mu_n - \mu_s \neq 0$ (there is a difference in average birth weight)

Step 2: Check conditions.

- *Independence.* The data come from a simple random sample, so observations are independent within and between groups.
- *Normality.* Both groups have large sample sizes ($n_n = 867$ and $n_s = 114$), both well above 30. Histograms of the data show no extreme outliers. Conditions are met.

Step 3: Compute the test statistic and p-value.

$$\bar{x}_n - \bar{x}_s = 7.27 - 6.68 = 0.59$$

$$SE = \sqrt{\frac{1.23^2}{867} + \frac{1.60^2}{114}} = \sqrt{\frac{1.5129}{867} + \frac{2.56}{114}} = \sqrt{0.001745 + 0.02246} = \sqrt{0.02420} = 0.156$$

$$T = \frac{0.59 - 0}{0.156} = 3.78$$

Using the conservative $df = \min(867 - 1, 114 - 1) = 113$, the one-tail area for $T = 3.78$ is approximately 0.000126. The two-tailed p-value is $2 \times 0.000126 \approx 0.00025$.

Step 4: Conclusion. With a p-value of 0.00025, which is far less than $\alpha = 0.05$, we reject H_0. The data provide strong evidence that there is a difference in average birth weight between babies born to smoking and non-smoking mothers. Specifically, babies born to non-smoking mothers weigh about 0.59 lbs more on average.

Try this in StatLens →

14.4.3 Worked example: Exam versions

An instructor created two versions of an exam and randomly distributed them. Sample statistics:

Table 14.2: Sample statistics for exam versions.

Version	n	Mean	SD	Min	Max
A	58	75.1	13.9	44	100
B	55	72.0	13.8	38	100

Test whether there is convincing evidence that one version is more difficult. Use $\alpha = 0.01$.

Step 1: State hypotheses.

- H_0: $\mu_A - \mu_B = 0$ (the exams are equally difficult)
- H_A: $\mu_A - \mu_B \neq 0$ (one exam is more difficult)

Step 2: Check conditions.

- *Independence.* Exams were randomly assigned, so independence is satisfied.
- *Normality.* Both groups have $n > 30$ and the sample statistics suggest roughly symmetric distributions. Conditions are met.

Step 3: Compute the test statistic and p-value.

$$\bar{x}_A - \bar{x}_B = 75.1 - 72.0 = 3.1$$

$$SE = \sqrt{\frac{13.9^2}{58} + \frac{13.8^2}{55}} = \sqrt{\frac{193.21}{58} + \frac{190.44}{55}} = \sqrt{3.331 + 3.463} = \sqrt{6.794} = 2.607$$

$$T = \frac{3.1}{2.607} = 1.19$$

Using $df = \min(57, 54) = 54$, the one-tail area for $T = 1.19$ is about 0.12. The two-tailed p-value is approximately 0.24.

Step 4: Conclusion. The p-value of 0.24 is much larger than $\alpha = 0.01$, so we fail to reject H_0. The data do not provide convincing evidence that one exam version is more difficult than the other.

Does the conclusion above mean that the two exams are equally difficult?

Show answer

No. Failing to reject H_0 does not prove that H_0 is true. The data are consistent with the exams being equally difficult, but they are also consistent with exam A being about 3.1 points easier. The test simply did not find strong enough evidence to distinguish these possibilities.

14.5 Paired t-test

14.5.1 Setup and hypotheses

When data are paired (see the confidence intervals for means chapter for what makes data "paired"), we compute the difference within each pair and test whether the mean difference is zero.

The test statistic for a paired mean difference is a T.

$$T = \frac{\bar{x}_{diff} - 0}{s_{diff}/\sqrt{n_{diff}}}$$

When the null hypothesis is true and conditions are met, T follows a t-distribution with $df = n_{diff} - 1$.

Conditions:

- **Independently sampled pairs.**
- **Normality of differences.** The differences should be approximately normal, or n_{diff} should be large. Check for extreme outliers in the differences.

Optional — see it by simulation. The paired test can also be run by simulation, as in the randomization tests chapter. Because the null says there is no difference, the two measurements in each pair are interchangeable, so we **flip the sign of each difference at random**, recompute the mean, and repeat to build a null distribution centered at zero. Open the Paired Randomization Test to try it. (Because it keeps the magnitudes of the differences, this is the sign-*flip* randomization test — not the lower-power nonparametric *sign test*, which counts only the signs.)

14.5.2 Worked example: Textbook prices

Is there evidence that UCLA Bookstore prices differ from Amazon prices, on average? Sample statistics for the 68 price differences (UCLA − Amazon):

Table 14.3: Sample statistics for the 68 price differences.

n_{diff}	$\bar{x}_{diff}$	s_{diff}
68	\$3.58	\$13.42

Step 1: State hypotheses.

- H_0: $\mu_{diff} = 0$ (no difference in average textbook price)
- H_A: $\mu_{diff} \neq 0$ (there is a difference in average price)

Step 2: Check conditions.

- *Independence.* The textbooks were a simple random sample of courses; pairs are independent.

- *Normality.* With $n_{diff} = 68$, the CLT applies. While some outliers exist (a few textbooks with large price differences), none are extreme enough to invalidate the approach given the large sample size.

Step 3: Compute the test statistic and p-value.

$$SE = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{13.42}{\sqrt{68}} = 1.628$$

$$T = \frac{3.58 - 0}{1.628} = 2.20$$

With $df = 68 - 1 = 67$, the one-tail area for $T = 2.20$ is approximately 0.0156. The two-tailed p-value is $2 \times 0.0156 = 0.0312$.

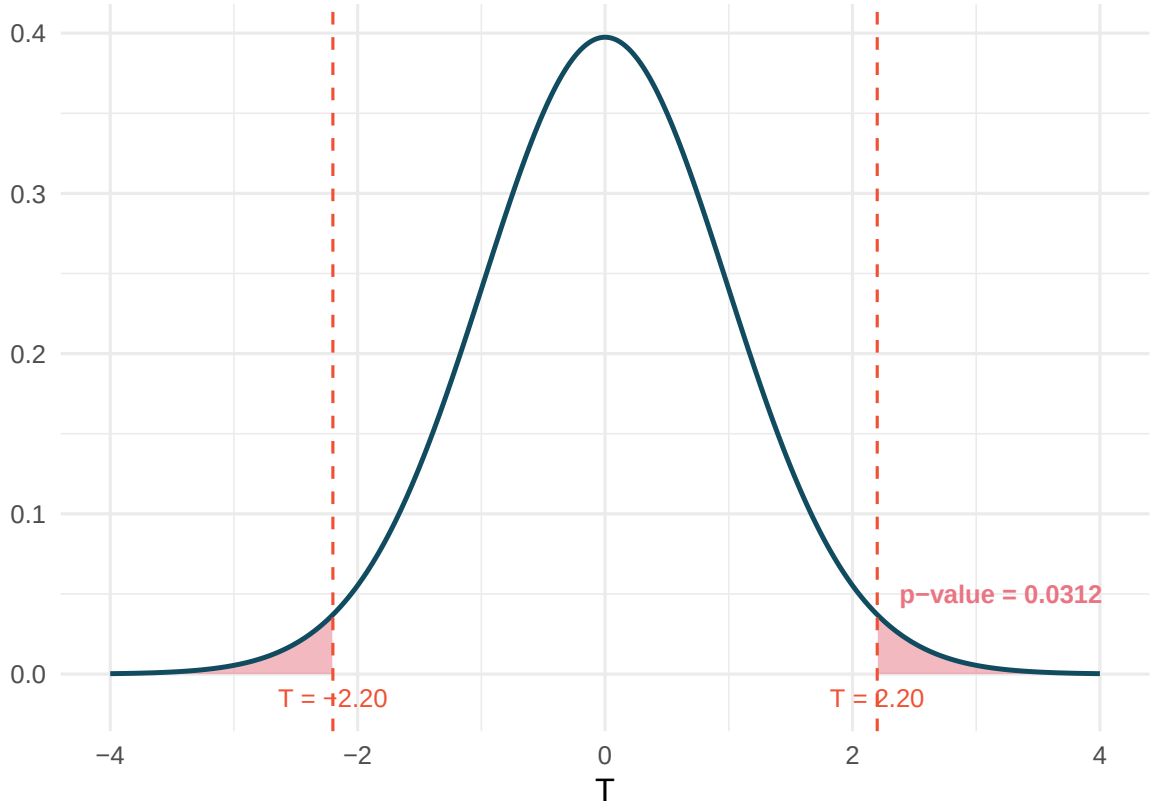


Figure 14.3: A t-distribution with df = 67 showing the two-sided p-value. The shaded areas in both tails represent the p-value of 0.0312. ↗ Try this live — drag the cutlines to reproduce this p-value with different df.

Step 4: Conclusion. The p-value of 0.0312 is less than $\alpha = 0.05$, so we reject H_0. The data provide evidence that UCLA Bookstore prices are different from Amazon prices on average. Based on the positive mean difference, the UCLA Bookstore tends to charge more.

Try this in StatLens →

Researchers measured tire tread (in cm) for 25 cars, each equipped with one Smooth Turn and one Quick Spin tire. The mean tread difference (Smooth Turn − Quick Spin) was $\bar{x}_{diff} = 0.002$ cm with

$s_{diff} = 0.004$ cm. Test whether there is a difference in average tread wear between the two brands. Use $\alpha = 0.05$.

Show answer

Hypotheses: $H_0 : \mu_{diff} = 0$, $H_A : \mu_{diff} \neq 0$. Conditions: cars are independent, $n = 25$ and no extreme outliers reported. $SE = 0.004/\sqrt{25} = 0.0008$. $T = 0.002/0.0008 = 2.50$. With $df = 24$, the two-tailed p-value is approximately 0.020. Since $0.020 < 0.05$, we reject H_0 and conclude there is evidence of a difference in tread wear between the two tire brands.

14.6 Connecting simulation-based and t-based tests

An important theme throughout this course is that simulation-based methods and mathematical models are complementary approaches to the same questions. When conditions are met, they should give similar answers. Here we compare the two approaches explicitly.

14.6.1 When they agree

For the birth weights example:

- **Randomization test:** Shuffle smoking/nonsmoking labels 10,000 times, compute the difference in means each time, and find the proportion of shuffled differences as extreme as 0.59 lbs. A difference that large essentially never appears (typically 0 of the 10,000 shuffles), so the p-value is below 0.0001.
- **t-test:** The p-value was 0.00025.

Both approaches lead to the same conclusion (reject H_0) by a wide margin. This agreement is typical when sample sizes are moderate to large and conditions are met.

14.6.2 When they might disagree

Disagreement arises primarily in two situations:

1. **Small samples from skewed populations.** The t-test assumes the sampling distribution is symmetric, which may not hold for small samples from skewed distributions. A simulation-based approach can capture the asymmetry.
2. **Extreme outliers.** The t-test relies on $\bar{x}$ and s, both of which are sensitive to outliers. A randomization test builds its null distribution by shuffling the actual data instead of assuming a t model, so its p-value does not depend on the normality assumption that outliers undermine (although the difference in means itself is still pulled around by outliers).

14.6.3 Practical guidance

- **When conditions are clearly met** (large samples, no extreme outliers), use the t-test – it is simpler and faster.
- **When conditions are borderline** (moderate samples, mild skewness), consider running both approaches. If they agree, report the t-test. If they disagree, investigate why and consider the simulation-based result more trustworthy.
- **When conditions are clearly violated** (small samples, strong skewness, extreme outliers), prefer the simulation-based approach.

14.7 Practical vs. statistical discernibility

Statistical discernibility $\neq$ practical importance.

A result is **statistically discernible** if the p-value is below the chosen discernibility level α. But statistical discernibility only means the observed effect is unlikely to be due to chance alone. It says nothing about whether the effect is *large enough to matter in practice*.

With very large samples, even trivially small differences can be statistically discernible. With very small samples, even large differences may not be statistically discernible (due to low power).

A large study of 50,000 students found that students who eat breakfast score an average of 0.5 points higher on a 100-point exam than students who skip breakfast ($p < 0.001$). Is this result practically important?

The result is statistically discernible (the p-value is very small), but the effect size of 0.5 points on a 100-point exam is tiny. A teacher or student would likely not consider half a point to be a meaningful difference. This is an example where statistical discernibility does not imply practical importance.

14.7.1 Effect sizes

To bridge the gap between statistical and practical importance, researchers often report **effect sizes** – standardized measures of how large the observed effect is.

Cohen's d is a common measure of effect size for comparing two means. It expresses the difference in means as a multiple of the pooled standard deviation:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

where s_p is the pooled standard deviation. General guidelines for interpreting d:

Table 14.4: Guidelines for interpreting Cohen's d.

$\lvert d \rvert$	Interpretation
0.2	Small effect
0.5	Medium effect
0.8	Large effect

These guidelines are rough benchmarks. What constitutes a "meaningful" effect depends on the context. In medicine, even a small effect size might be important if it concerns a life-threatening condition. In education, a medium effect might be required for a costly intervention to be worthwhile.

In the birth weights study, the difference was 0.59 lbs with a pooled standard deviation of approximately 1.28 lbs. Compute Cohen's d and interpret it.

Show answer

$d = 0.59/1.28 = 0.46$. This is between a small and medium effect. While the difference is statistically discernible, the practical impact is moderate – about half a standard deviation in birth weight.

14.8 Chapter review

14.8.1 Summary

This chapter presented hypothesis tests for means using the t-distribution:

- **One-sample t-test**: Tests whether a population mean μ equals a hypothesized value μ_0. Test statistic: $T = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ with $df = n - 1$.
- **Two-sample t-test**: Tests whether two population means are equal. Test statistic: $T = \frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{s_1^2/n_1+s_2^2/n_2}}$ with $df = \min(n_1 - 1, n_2 - 1)$.
- **Paired t-test**: Tests whether the mean of paired differences equals zero. Compute differences first, then apply the one-sample t-test to the differences: $T = \frac{\bar{x}_{diff}}{s_{diff}/\sqrt{n_{diff}}}$ with $df = n_{diff} - 1$.
- All three tests require **independence** and **approximate normality or large sample size**.
- **Simulation-based and t-based tests** generally agree when conditions are met. When conditions are questionable, simulation is more reliable.
- **Statistical discernibility** (small p-value) does not imply **practical importance** (large enough to matter). Effect sizes like **Cohen's d** help assess practical importance.

14.8.2 Key terms

T test statistic, one-sample t-test, two-sample t-test, paired t-test, degrees of freedom, p-value, discernibility level (α), statistical discernibility, practical importance, effect size, Cohen's d.

14.9 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Find the p-value, I.** A random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given sample size and test statistic. Also determine if the null hypothesis would be rejected at $\alpha = 0.05$.

   a. $n = 11, T = 1.91$
   b. $n = 17, T = -3.45$
   c. $n = 7, T = 0.83$
   d. $n = 28, T = 2.13$

2. **Find the p-value, II.** A random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given sample size and test statistic. Also determine if the null hypothesis would be rejected at $\alpha = 0.01$.

   a. $n = 26, T = 2.485$
   b. $n = 18, T = 0.5$

3. **Length of gestation, hypothesis test.** In this exercise we work with a random sample of 1,000 cases from the dataset released by the United States Department of Health and Human Services in 2014. Provided below are sample statistics for gestation (length of pregnancy, measured in weeks) of births in this sample.

Min	Q1	Median	Mean	Q3	Max	SD	IQR
21	38	39	38.7	40	46	2.6	2

a. What is the point estimate for the average length of pregnancy for all women? What about the median?

b. You might have heard that human gestation is typically 40 weeks. Using the data, perform a complete hypothesis test, using mathematical models, to assess the 40 week claim. State the null and alternative hypotheses, find the T score, find the p-value, and provide a conclusion in context of the data.

c. A quick internet search validates the claim of "40 weeks gestation" for humans. A friend of yours claims that there are different ways to measure gestation (starting at first day of last period, ovulation, or conception) which will result in estimates that are a week or two different. Another friend mentions that recent increases in cesarean births is likely to have decreased length of gestation. Do the data provide a mechanism to distinguish between your two friends' claims?

4. **Interpreting p-values for population mean.** For each of the following statements, indicate if they are a true or false interpretation of the p-value. If false, provide a reason or correction to the misinterpretation. You are wondering if the average amount of cereal in a 10oz cereal box is greater than 10oz. You collect 50 boxes of cereal, weigh them carefully, find a T score, and a p-value of 0.23.

a. The probability that the average weight of all cereal boxes is 10 oz is 0.23.

b. The probability that the average weight of all cereal boxes is greater than 10 oz is 0.23.

c. Because the p-value is 0.23, the average weight of all cereal boxes is 10 oz.

d. Because the p-value is small, the population average must be just barely above 10 oz.

e. If H_0 is true, the probability of observing another sample with an average as or more extreme as the data is 0.23.

5. **Sleep habits of New Yorkers.** New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. The point estimate suggests New Yorkers sleep less than 8 hours a night on average. Evaluate the claim that New York is the city that never sleeps keeping in mind that, despite this claim, the true average number of hours New Yorkers sleep could be less than 8 hours or more than 8 hours.

n	Mean	SD	Min	Max
25	7.73	0.77	6.17	9.78

a. Write the hypotheses in symbols and in words.

b. Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.

c. Find and interpret the p-value in this context. Drawing a picture may be helpful.

d. What is the conclusion of the hypothesis test?

e. If you were to construct a 90% confidence interval that corresponded to this hypothesis test, would you expect 8 hours to be in the interval?

6. **Find the mean.** You are given the hypotheses shown below. We know that the sample standard deviation is 8 and the sample size is 20. For what sample mean would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

$$H_0 : \mu = 60 \qquad H_A : \mu \neq 60$$

7. **Play the piano.** Georgianna claims that in a small city renowned for its music school, the average child takes less than 5 years of piano lessons. We have a random sample of 20 children from the city, with a mean of 4.6 years of piano lessons and a standard deviation of 2.2 years.

a. Evaluate Georgianna's claim (or that the opposite might be true) using a hypothesis test.

b. Construct a 95% confidence interval for the number of years students in this city take piano lessons, and interpret it in context of the data.

c. Do your results from the hypothesis test and the confidence interval agree? Explain your reasoning.

8. **Auto exhaust and lead exposure.** Researchers interested in lead exposure due to car exhaust sampled the blood of 52 police officers subjected to constant inhalation of automobile exhaust fumes while working traffic enforcement in a primarily urban environment. The blood samples of these officers had an average lead concentration of 124.32 μg/l and a SD of 37.74 μg/l; a previous study of individuals from a nearby suburb, with no history of exposure, found an average blood level concentration of 35 μg/l. (Mortada et al. 2000)

a. Write down the hypotheses that would be appropriate for testing if the police officers appear to have been exposed to a different concentration of lead.

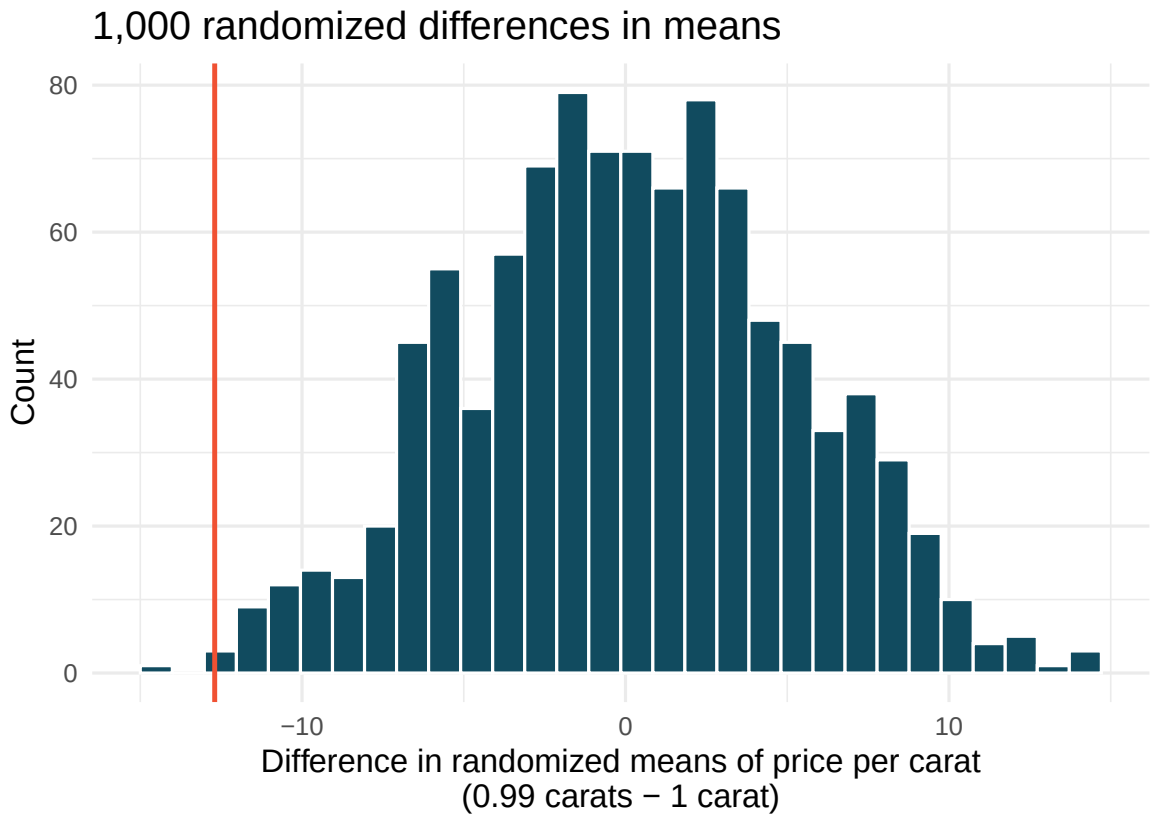
b. Explicitly state and check all conditions necessary for inference on these data.

c. Test the hypothesis that the downtown police officers have a higher lead exposure than the group in the previous study. Interpret your results in context.

9. **Experimental baker.** A baker working on perfecting their bagel recipe is experimenting with active dry (AD) and instant (I) yeast. They bake a dozen bagels with each type of yeast and score each bagel on a scale of 1 to 10 on how well the bagels rise. They come up with the following set of hypotheses for evaluating whether there is a difference in the average rise of bagels baked with active dry and instant yeast. What is wrong with the hypotheses as stated?

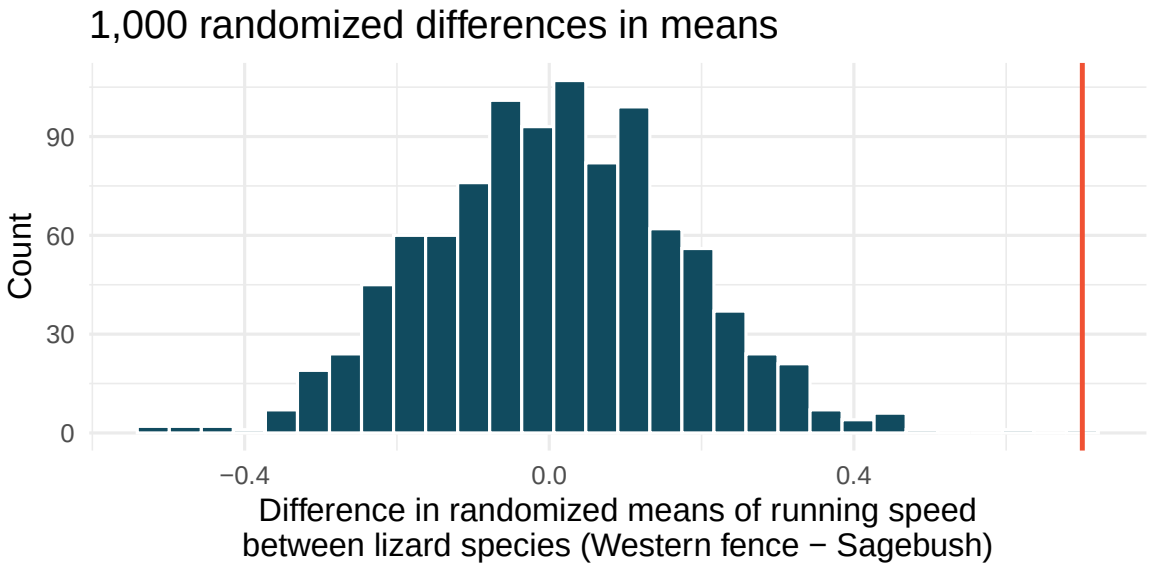
$$H_0 : \bar{x}_{AD} \leq \bar{x}_I \qquad H_A : \bar{x}_{AD} > \bar{x}_I$$

10. **Diamonds, randomization test.** The prices of diamonds go up as the carat weight increases, but the increase is not smooth. For example, the difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but the price of a 1 carat diamond tends to be much higher than the price of a 0.99 carat diamond. We have two random samples of diamonds: 23 0.99 carat diamonds and 23 1 carat diamonds. In order to be able to compare equivalent units, we first divide the price for each diamond by 100 times its weight in carats. That is, for a 0.99 carat diamond, we divide the price by 99 and for a 1 carat diamond, we divide it by 100. Then, we randomize the carat weight to the price values in order simulate the null distribution of differences in average prices of 0.99 carat and 1 carat diamonds. The null distribution (with 1,000 randomized differences) is shown below and depicts the distribution of differences in sample means (of price per carat) if there really was no difference in the population from which these diamonds came. (Wickham 2016)



Using the randomization distribution, conduct a hypothesis test to evaluate if there is a difference between the prices per carat of diamonds that weigh 0.99 carats and diamonds that weigh 1 carat. Make sure to state your hypotheses clearly and interpret your results in context of the data. (Wickham 2016)

11. **Lizards running, randomization test.** In order to assess physiological characteristics of common lizards, data on top speeds (in m/sec) measured on a laboratory race track for two species of lizards: Western fence lizard (Sceloporus occidentalis) and Sagebrush lizard (Sceloporus graciosus). The original observed difference in lizard speeds is $\bar{x}_{Westernfence} - \bar{x}_{Sagebrush} = 0.7$m/sec. The histogram below shows the distribution of average differences when speed has been randomly allocated across lizard species 1,000 times. Using the randomization distribution, conduct a hypothesis test to evaluate if there is a difference between the average speed of the Western fence lizard as compared to the Sagebrush lizard. Make sure to state your hypotheses clearly and interpret your results in context of the data. (Adolph 1987)



12. **Possible randomized means.** Data were collected on data from two groups (A and B). There were 3 measurements taken on Group A and two measurements in Group B.

Group	Measurement 1	Measurement 2	Measurement 3
A	1	15	5
B	7	3	NA

If the data are (repeatedly) randomly allocated across the two conditions, provide the following: (1) the values which are assigned to group A, (2) the values which are assigned to group B, and (3) the difference in averages $(\bar{x}_A - \bar{x}_B)$ for each of the following:

a. When the randomized difference in averages is as large as possible.

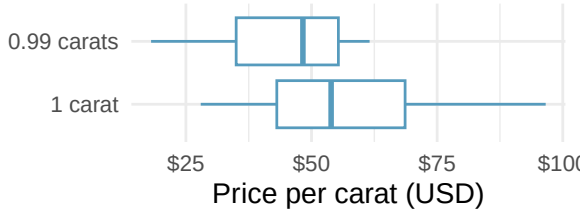
b. When the randomized difference in averages is as small as possible (a big in magnitude negative number).

c. When the randomized difference in averages is as close to zero as possible.

d. When the observed values are randomly assigned to the two groups, to which of the previous parts would you expect the difference in means to fall closest? Explain your reasoning.

13. **Diamonds, mathematical test.** We have data on two random samples of diamonds: one with diamonds that weigh 0.99 carats and one with diamonds that weigh 1 carat. Each sample has 23 diamonds. Sample statistics for the price per carat of diamonds in each sample are provided below. Conduct a hypothesis test using a mathematical model to evaluate if there is a difference between the prices per carat of diamonds that weigh 0.99 carats and diamonds that weigh 1 carat Make sure to state your hypotheses clearly, check relevant conditions, and interpret your results in context of the data. (Wickham 2016)

	Mean	SD	n
0.99 carats	\$44.51	\$13.32	23
1 carat	\$57.20	\$18.19	23



14. **A/B testing.** A/B testing is a user experience research methodology where two variants of a page are shown to users at random. A company wants to evaluate whether users will spend more time, on average, on Page A or Page B using an A/B test. Two user experience designers at the company, Lucie and Müge, are tasked with conducting the analysis of the data collected. They agree on how the null hypothesis should be set: on average, users spend the same amount of time on Page A and Page B. Lucie believes that Page B will provide a better experience for users and hence wants to use a one-tailed test, Müge believes that a two-tailed test would be a better choice. Which designer do you agree with, and why?

15. **Mindfulness intervention for nurses.** In order to address extremely challenging and stressful situations for intensive care unit nurses, researchers ran a mindfulness-based intervention (MBI) study on 60 nurses working in three hospitals in El-Beheira, Egypt. The participants were randomly allocated to one of the two groups: the treatment group (MBI) received 8 MBI sessions and the control group received no intervention. The nurses' emotional exhaustion was measured using 9 items from a questionnaire of the Maslach Burnout Inventory-Human Services Survey for Medical Personnel; the questions are recorded on a Likert scale where 0 indicated "Never" and 6 indicates "Every day". Nurses in the treatment group had an emotional exhaustion score of 15.47, with a standard deviation of 4.44, and nurses in the control group had an emotional exhaustion score of 32.43, with a standard deviation of 8.87. Do these data provide convincing evidence that the emotional exhaustion decrease is different for the patients in the treatment group compared to the control group? Assume that conditions for conducting inference using mathematical models are satisfied. (Othman, Hassan, and Mohamed 2023)

16. **Chicken diet: horsebean vs. linseed.** Chicken farming is a multi-billion dollar industry, and any methods that increase the growth rate of young chicks can reduce consumer costs while increasing company profits, possibly by millions of dollars. An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. In this exercise we consider chicks that were fed horsebean and linseed. Below are some summary statistics from this dataset along with box plots showing the distribution of weights by feed type. (McNeil 1977)

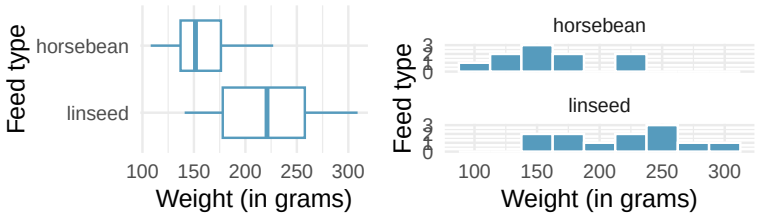
    a. Describe the distributions of weights of chickens that were fed horsebean and linseed.

    b. Do these data provide strong evidence that the average weights of chickens that were fed linseed and horsebean are different? Use a 5% discernibility level.

    c. What type of error might we have committed? Explain.

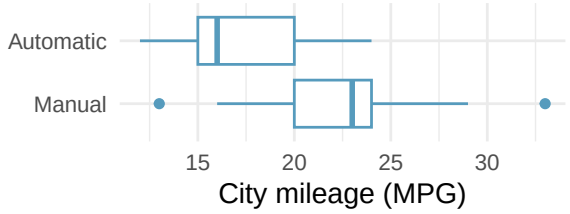
    d. Would your conclusion change if we used $\alpha = 0.01$?

	Horsebean	Linseed
Mean	160.2	218.8
SD	38.6	52.2
n	10.0	12.0



17. **Fuel efficiency in the city.** Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions manufactured in 2021. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? (US DOE EPA 2021)

CITY	Mean	SD	n
Automatic	17.4	3.44	25
Manual	22.7	4.58	25

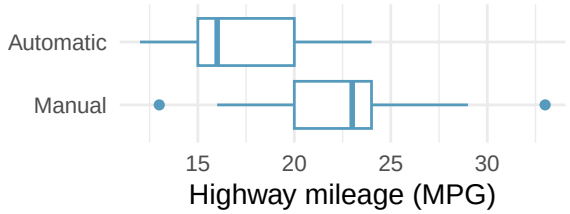


18. **Chicken diet: casein vs. soybean.** Casein is a common weight gain supplement for humans. Does it have an effect on chickens? An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. In this exercise we consider chicks that were fed casein and soybean. Assume that the conditions for conducting inference using mathematical models are met, and using the data provided below, test the hypothesis that the average weight of chickens that were fed casein is different than the average weight of chickens that were fed soybean. If your hypothesis test yields a statistically discernible result, discuss whether the higher average weight of chickens can be attributed to the casein diet. (McNeil 1977)

Feed type	Mean	SD	n
casein	323.58	64.43	12
soybean	246.43	54.13	14

19. **Fuel efficiency on the highway.** Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions manufactured in 2021. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average highway mileage? (US DOE EPA 2021)

HIGHWAY	Mean	SD	n
Automatic	23.7	3.90	25
Manual	30.9	5.13	25

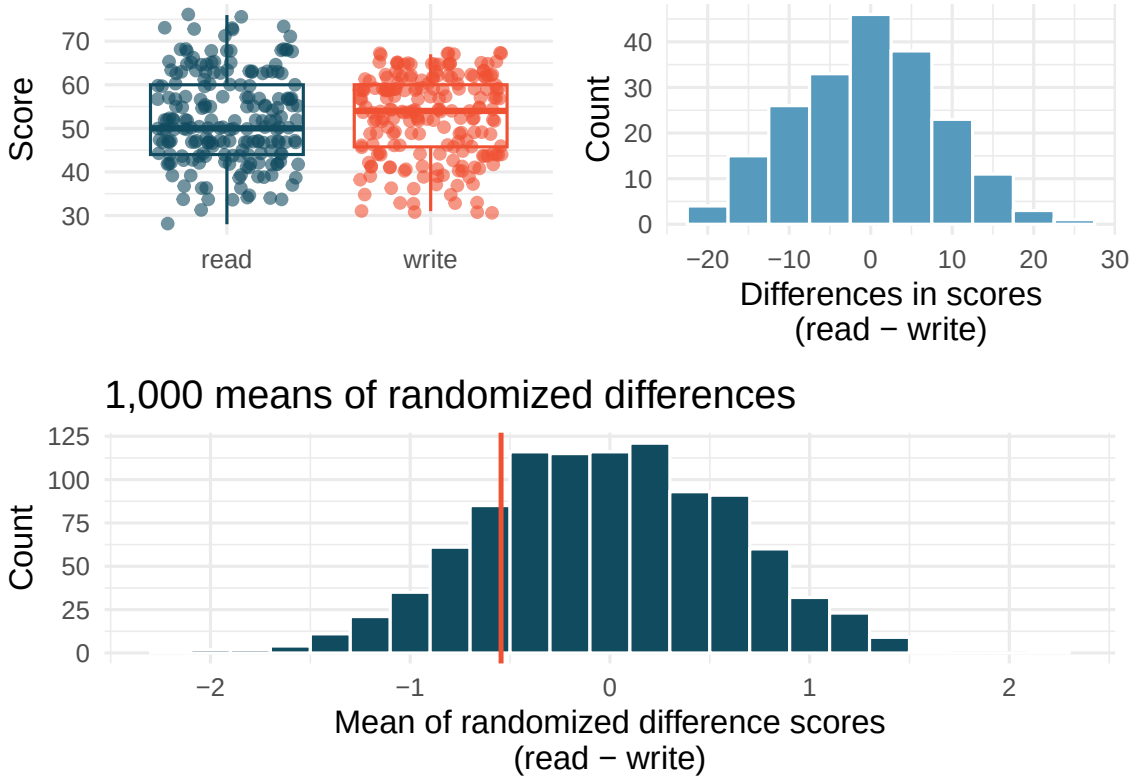


20. **Gaming, distracted eating, and intake.** A group of researchers who are interested in the possible effects of distracting stimuli during eating, such as an increase or decrease in the amount of food consumption, monitored food intake for a group of 44 patients who were randomized into two equal groups. The treatment group ate lunch while playing solitaire, and the control group ate lunch without any added distractions. Patients in the treatment group ate 52.1 grams of biscuits, with a standard deviation of 45.1 grams, and patients in the control group ate 27.1 grams of biscuits, with a standard deviation of 26.4 grams. Do these data provide convincing evidence that the average food intake (measured in amount of biscuits consumed) is different for the patients in the treatment group compared to the control group? Assume that conditions for conducting inference using mathematical models are satisfied. (Oldham-Cooper et al. 2011)

21. **Gaming, distracted eating, and recall.** A group of researchers who are interested in the possible effects of distracting stimuli during eating, such as an increase or decrease in the amount of food consumption, monitored food intake for a group of 44 patients who were randomized into two equal groups. The 22 patients in the treatment group who ate their lunch while playing solitaire were asked to do a serial-order recall of the food lunch items they ate. The average number of items recalled by the patients in this group was 4. 9, with a standard deviation of 1.8. The average number of items recalled by the patients in the control group (no distraction) was 6.1, with a standard deviation of 1.8. Do these data provide strong evidence that the average numbers of food items recalled by the patients in the treatment and control groups are different? Assume that conditions for conducting inference using mathematical models are satisfied. (Oldham-Cooper et al. 2011)

22. **High School and Beyond, randomization test.** The National Center of Education Statistics conducted a survey of high school seniors, collecting test data on reading, writing, and several other subjects. Here we examine a simple random sample of 200 students from this survey.

Side-by-side box plots of reading and writing scores as well as a histogram of the differences in scores are shown below. Also provided below is a histogram of randomized averages of paired differences of scores (read - write), with the observed difference ($\bar{x}_{read-write} = -0.545$) marked with a red vertical line. The randomization distribution was produced by doing the following 1000 times: for each student, the two scores were randomly assigned to either read or write, and the average was taken across all students in the sample.



a. Is there a clear difference in the average reading and writing scores?

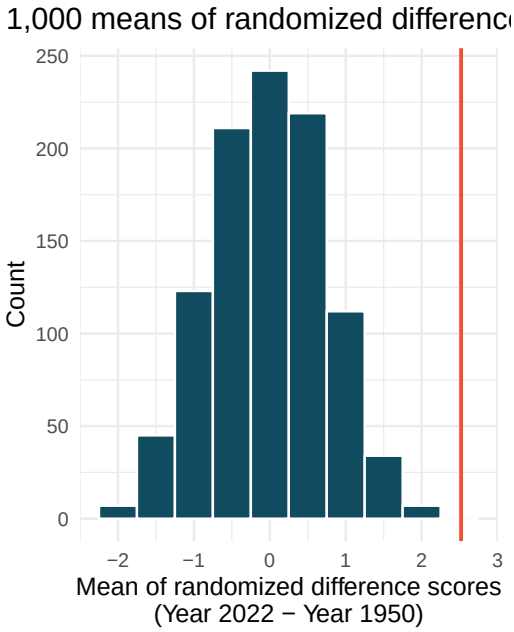
b. Are the reading and writing scores of each student independent of each other?

c. Create hypotheses appropriate for the following research question: is there an evident difference in the average scores of students in the reading and writing exam?

d. Is the average of the observed difference in scores ($\bar{x}_{read-write} = -0.545$) consistent with the distribution of randomized average differences? Explain.

e. Do these data provide convincing evidence of a difference between the average scores on the two exams? Estimate the p-value from the randomization test, and conclude the hypothesis test using words like "score on reading test" and "score on writing test."

23. **Global warming, randomization test.** Let's consider a limited set of climate data, examining temperature differences in 1950 vs 2022. We sampled 26 locations in the US from the National Oceanic and Atmospheric Administration's (NOAA) historical data, where the data was available for both years of interest. (NOAA 2023) The data are not a random sample, but they are selected to be a representative sample across the land area of the lower 48 United States. Using the hottest day of the year as a measure can make the results susceptible to outliers. Instead, to get a sense for how hot a year was, we calculate the 90th percentile; that is, we find the maximum temperature on the day that was hotter than 90% of the days that year. We want to know: is the 90th percentile high temperature greater in 2022 or in 1950? The difference in 90th percentile high temperature (high temperature for 2022 - high temperature for 1950) was calculated for each of the 26 locations. The average of the 26 differences was 2.52°F with a standard deviation of 2.95°F. We are interested in determining whether these data provide strong evidence that the 90th percentile high temperature is higher in 2022 than in 1950.



a. Create hypotheses appropriate for the following research question: is there an evident difference in the 90th percentile high temp across the two years (1950 and 2022)?

b. Is the average of the observed difference in scores ($\bar{x}_{2022-1950} = 2.53$^∘$F$) consistent with the distribution of randomized average differences? Explain.

c. Do these data provide convincing evidence of a difference between the 90th percentile high temperature? Estimate the p-value from the randomization test, and conclude the hypothesis test using words like "90th percentile high temperature in 1950" and "90th percentile high temperature in 2022."

24. **High School and Beyond, mathematical test.** We considered the differences between the reading and writing scores of a random sample of 200 students who took the High School and Beyond Survey.

a. Create hypotheses appropriate for the following research question: is there an evident difference in the average scores of students in the reading and writing exam?

b. Check the conditions required to complete this test.

c. The average observed difference in scores is $\bar{x}_{read-write} = -0.545$, and the standard deviation of the differences is $s_{read-write} = 8.887$ points. Do these data provide convincing evidence of a difference between the average scores on the two exams?

d. What type of error might we have made? Explain what the error means in the context of the application.

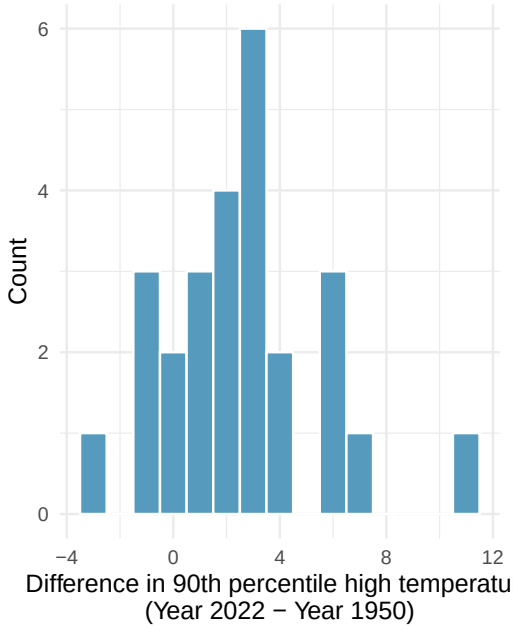
e. Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

25. **Global warming, mathematical test.** We considered the change in the 90th percentile high temperature in 1950 versus 2022 at 26 sampled locations from the NOAA database. (NOAA 2023) The mean and standard deviation of the reported differences are 2.53°F and 2.95°F.

    a. Is there a relationship between the observations collected in 1950 and 2022? Or are the observations in the two groups independent? Explain your reasoning.

    b. Write hypotheses for this research in symbols and in words.

    c. Check the conditions required to complete this test.



    d. Calculate the test statistic and find the p-value.

    e. Use $\alpha = 0.05$ to evaluate the test, and interpret your conclusion in context.

    f. What type of error might we have made? Explain in context what the error means.

    g. Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the 90th percentile high temperture from 1950 to 2022 to include 0? Explain your reasoning.

26. **Possible paired randomized differences.** Two observations were collected on each of five people. Which of the following could be a possible randomization of the paired differences given in the table below? If the set of values could not be a randomized set of differences, indicate why not.

    a. -2, 1, 1, 11, -2
    b. -4, 11, -2, 0, 1
    c. -2, 2, -11, 11, -2, 2, 0, 1, -1
    d. 0, -1, 2, -4, 11
    e. 4, -11, 2, 0, -1

	People				
	1	2	3	4	5
Observation 1	3	14	4	5	10
Observation 2	7	3	6	5	9
Difference	-4	11	-2	0	1

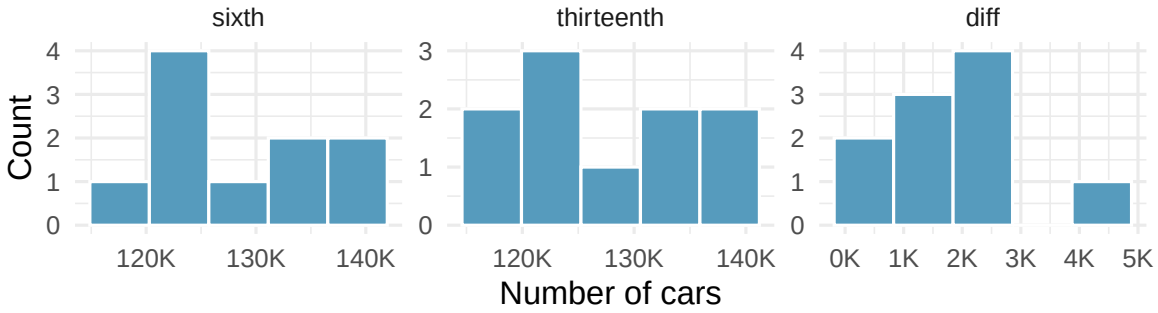
27. **Study environment.** In order to test the effects of listening to music while studying versus studying in silence, students agree to be randomized to two treatments (i.e., study with music or study in silence). There are two exams during the semester, so the researchers can either randomize the students to have one exam with music and one with silence (randomly selecting which exam corresponds to which study environment) or the researchers can randomize the students to one study habit for both exams.

The researchers are interested in estimating the true population difference of exam score for those who listen to music while studying as compared to those who study in silence.

a. Describe the experiment which is consistent with a paired designed experiment. How is the treatment assigned, and how are the data collected such that the observations are paired?

b. Describe the experiment which is consistent with an indpenedent samples experiment. How is the treatment assigned, and how are the data collected such that the observations are independent?

28. **Friday the 13th, traffic.** In the early 1990's, researchers in the UK collected data on traffic flow on Friday the 13th with the goal of addressing issues of how superstitions regarding Friday the 13th affect human behavior and and whether Friday the 13th is an unlucky day. The histograms below show the distributions of numbers of cars passing by a specific intersection on Friday the 6th and Friday the 13th for many such date pairs. Also provided are some sample statistics, where the difference is the number of cars on the 6th minus the number of cars on the 13th. (Scanlon et al. 1993)



	n	Mean	SD
sixth	10	128,385	7,259
thirteenth	10	126,550	7,664
diff	10	1,836	1,176

a. Are there any underlying structures in these data that should be considered in an analysis? Explain.

b. What are the hypotheses for evaluating whether the number of people out on Friday the 6th is different than the number out on Friday the 13th?

c. Check conditions to carry out the hypothesis test from part (b) using mathematical models.

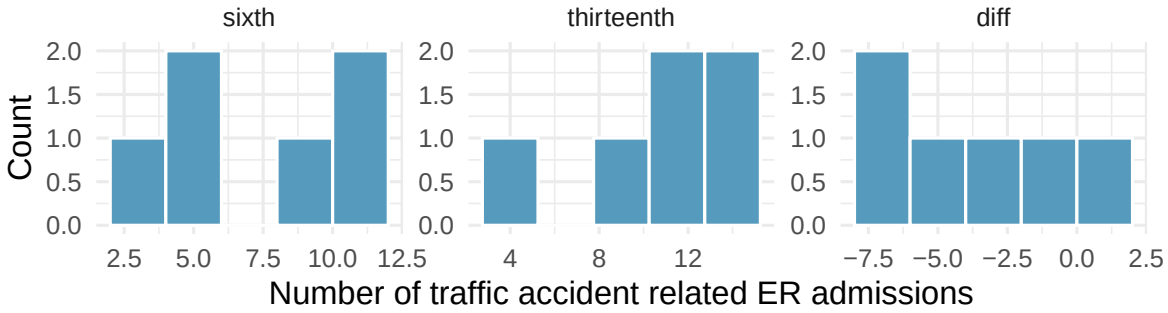
d. Calculate the test statistic and the p-value.

e. What is the conclusion of the hypothesis test?

f. Interpret the p-value in this context.

g. What type of error might have been made in the conclusion of your test? Explain.

29. **Friday the 13th, accidents.** In the early 1990's, researchers in the UK collected data the number of traffic accident related emergency room (ER) admissions on Friday the 13th with the goal of addressing issues of how superstitions regarding Friday the 13th affect human behavior and and whether Friday the 13th is an unlucky day. The histograms below show the distributions of numbers of ER admissions at specific emergency rooms on Friday the 6th and Friday the 13th for many such date pairs. Also provided are some sample statistics, where the difference is the ER admissions on the 6th minus the ER admissions on the 13th.(Scanlon et al. 1993)



	n	Mean	SD
sixth	6	8	3
thirteenth	6	11	4
diff	6	-3	3

a. Conduct a hypothesis test using mathematical models to evaluate if there is a difference between the average numbers of traffic accident related emergency room admissions between Friday the 6th and Friday the 13th.

b. Calculate a 95% confidence interval using mathematical models for the difference between the average numbers of traffic accident related emergency room admissions between Friday the 6th and Friday the 13th.

c. The conclusion of the original study states, "Friday 13th is unlucky for some. The risk of hospital admission as a result of a transport accident may be increased by as much as 52%. Staying at home is recommended." Do you agree with this statement? Explain your reasoning.

Datasets: `births14` (openintro) | `diamonds` (ggplot2) | `lizard_run` (openintro) | `chickwts` (datasets) | `epa2021` (openintro) | `hsb2` (openintro) | `us_temperature` (openintro) | `friday` (openintro)

StatLens Exercises

These exercises focus on the parts of inference a calculator *can't* do for you: **framing the question**, **choosing the right procedure**, **checking conditions**, and **interpreting** what a tool reports. Where a link is given, let StatLens do the arithmetic and spend your effort on the reasoning.

1. **Frame the question.** A manufacturer advertises that its energy bars contain a mean of 220 calories. A consumer group buys a random sample of $n = 36$ bars and measures the actual calorie content of each.

    a. Identify the parameter of interest in words and give its symbol.

   b. Write the null and alternative hypotheses for a test of whether the *true* mean differs from 220. Is your test one-sided or two-sided, and why?
   c. Suppose the consumer group instead suspected the manufacturer was *under-filling* the bars. Rewrite the alternative hypothesis. What is the practical danger of switching from a two-sided to a one-sided test *after* looking at the data?

2. **Choose the procedure.** For each scenario, name the procedure — **one-sample** t, **two-sample** t (independent groups), **paired** t, or a procedure **beyond this chapter** (more than two means, or a proportion).

   a. Each of 20 patients has their blood pressure measured before and after a meditation program; ask whether the program lowers blood pressure on average.
   b. An agronomist compares the mean yield (bushels/acre) of two corn varieties planted in 30 randomly chosen plots each.
   c. Is the proportion of voters who support a candidate more than 50%?
   d. Test whether the mean fuel economy of a single car model differs from the EPA-rated 32 mpg.
   e. Compare the mean weight gain of mice on three different diets.

3. **Check conditions before you trust the test.** Open the One-Sample t tool with any small numeric dataset (or paste in $n = 15$ values of your own).

   a. What two conditions does the one-sample t-test require? State each in plain language.
   b. The tool displays a dotplot or histogram of the data alongside the t-statistic. With $n = 15$, what would you look for in that picture to decide whether the test is trustworthy?
   c. Suppose the data show an extreme outlier — one observation more than three SDs from the rest. Should you still run the t-test? What two questions should you ask before deciding what to do with the outlier?

4. **Run it and interpret it (two means).** Open the Two-Sample t tool with any two-group numeric dataset (or use the bundled examples).

   a. Identify the explanatory and response variables, and define $\mu_1 - \mu_2$ in words.
   b. Run the test. Report the test statistic, degrees of freedom, and p-value. State your decision at $\alpha = 0.05$.
   c. Write a one-sentence conclusion **in context**. Then write *another* sentence that translates the p-value into plain language — without using the words "probability that H_0 is true."
   d. Which statement is the correct interpretation of the p-value?
      i. The probability that the null hypothesis is true given the data.
      ii. The probability of observing a difference at least as extreme as the one we got, *if* H_0 were true.
      iii. The probability that $\mu_1 = \mu_2$.

5. **Simulation vs. analytic — when do they agree?** A randomization test and a two-sample t-test are two routes to the same conclusion about whether two group means differ. Sometimes they land in nearly the same place; sometimes they don't.

   First a well-behaved case. Open any moderate-n two-group numeric dataset in the Two-Sample t tool and record the p-value, then open the *same* data in the Randomization Test for the Difference in Means and run 1000 shuffles.

   a. How close are the two p-values?

   Now construct (or pick) a trouble case — a small dataset ($n_1, n_2 \leq 10$ each) where one group has a visible outlier or a strong skew. Run *both* procedures again.

b. Are the p-values still close? Where do they diverge — is the t-test's p-value smaller or larger than the randomization p-value? In which direction?
c. Explain the pattern. Why do the two procedures agree on well-behaved data but diverge in the trouble case? Tie your explanation to (i) the role of normality in the t distribution and (ii) what the randomization test *does* assume (only exchangeability under H_0).
d. A student says, "The randomization test gave a bigger p-value, so the t-test must be wrong." Why is that the wrong framing? Which test should you trust more when they disagree, and why?

Compute — Unit Review

Unit 3 revisits the same questions you answered by simulation — *is there an effect?* and *how big is it?* — but answers them with **formulas**. When the conditions hold, the simulated distributions of Unit 2 are approximately **normal** (or **t**), so a short calculation replaces thousands of resamples. This review consolidates *which* formula goes with which question, and *when you're allowed to use it*.

The big picture

Mathematical models showed that, under the right conditions, the sampling distribution of a statistic is approximately **normal**. That single fact powers the rest of the part:

- **Inference for proportions** — the normal model gives the **z**-procedures for one and two proportions.
- **Confidence intervals for means** and **hypothesis tests for means** — because we estimate the standard deviation from the data, means use the **t**-distribution (one-sample, two-sample, and paired).

The logic is identical to Unit 2 — a **test** gives a p-value, an **interval** gives a range — but the null/sampling distribution now comes from a *model* instead of a simulation. The trade: formulas are fast, but they're only valid when their **conditions** are met.

Which z- or t-procedure?

Pin the procedure with three questions — **proportion or mean? one group or two? (for means) paired or independent?** — then check the condition.

Situation	Procedure	Key condition
One **proportion** vs. a value	one-proportion **z**	$np_0 \geq 10$ and $n(1-p_0) \geq 10$
Two **proportions** compared	two-proportion **z**	each group has ≥10 successes & ≥10 failures
One **mean** vs. a value	one-sample **t**	roughly normal, or large n
Two **means**, independent groups	two-sample **t**	each group roughly normal / large
Paired measurements (before/after, matched)	paired **t** (t on the differences)	the *differences* roughly normal

Paired vs. independent is a design question, not a math question. If each value in one group is naturally linked to a specific value in the other (same subject twice, matched pairs), it's **paired** — analyze the differences. If the two groups are separate samples, it's **two-sample**.

Simulation or formula?

Both routes answer the same question; pick by **conditions**:

1. **Identify the procedure** from the table (proportion/mean, one/two, paired).
2. **Check its condition.**
   - **Conditions met** → the **z/t formula** is valid and convenient.
   - **Conditions fail** (e.g. too few successes/failures, small skewed sample) → fall back to the **simulation** method from Unit 2 (randomization test or bootstrap), which makes no normality assumption.
3. Remember they **agree when both apply** — the formula is the shortcut the simulation justifies.

Key ideas to carry forward

Sampling distribution · **standard error** · **z-procedure** (proportions) · **t-distribution** (means) · **degrees of freedom** · **success–failure condition** · **pooled proportion** · **paired vs. two-sample** · **margin of error** · **critical value** (z^*, t^*) · **p-value** · **confidence level**.

Common pitfalls

- **Using a z-test on a proportion when** $np_0 < 10$ — the normal model is unreliable; use simulation (this is the malaria/medical-consultant lesson).
- **Treating paired data as two independent samples** (or vice versa) — wrong standard error, wrong df.
- **"Fail to reject" ≠ "accept the null."**
- **A bigger sample shrinks the margin of error and raises power** — but never fixes bias.
- **A confidence interval is about the procedure's long-run capture rate**, not a probability for the one interval.

Exercises

The Unit 3 skill is **naming the right z/t procedure and checking whether you're allowed to use it.** Answers are provided so you can check your work.

1. **Name the procedure.** For each scenario, give the specific analytic procedure (one-proportion z, two-proportion z, one-sample t, two-sample t, or paired t).

   a. Is the proportion of defective chips different from the 1% target?
   b. Do graduates of program A earn a higher *average* starting salary than graduates of program B (two separate samples)?
   c. Did blood pressure change after a 12-week program, measured on the *same* patients before and after?
   d. Is the average dosage in a batch different from the labeled 500 mg?
   e. Is the proportion of voters favoring a measure different between two cities?

   Show answer

   a. one-proportion z (vs. 0.01). **b.** two-sample t. **c.** paired t (test the before–after differences). **d.** one-sample t (vs. 500). **e.** two-proportion z.

2. **Are you allowed to use it?** A clinic claims its complication rate is below the national 10%. In a sample of $n = 62$ patients, 3 had complications. Should you trust a one-proportion **z**-test here? What should you do instead?

Show answer

Check the condition with $p_0 = 0.10$: $np_0 = 62(0.10) = 6.2 < 10$, so the **success–failure condition fails** — the z-test is **not** trustworthy. Use a **simulation** (a bootstrap interval or randomization test), which makes no normality assumption. ("Run the test" sometimes correctly answers "the conditions don't support this test.")

3. **Paired or two-sample?** Two studies compare a drug to a placebo. Study X gives the drug to one group and placebo to a *different* group. Study Y gives *each* subject the drug for a month and placebo for a month and compares within subject. Which t-procedure does each use?

   Show answer

   Study X → two-sample t (independent groups). **Study Y → paired t** (each subject is their own control; analyze the within-subject differences). Same comparison, different design → different procedure and standard error.

4. **Simulation vs. formula.** On a tidy dataset with 200 observations and dozens of successes and failures in each group, a two-proportion z-test and a randomization test give nearly the same p-value. On a 20-observation dataset with a zero cell, they diverge. Explain the pattern.

   Show answer

   With large counts the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is close to normal, so the **formula and the simulation agree**. With tiny, lopsided counts the normal model is a poor fit, so the z-test's p-value is unreliable and **diverges** from the simulation — which stays valid. Divergence is a *signal that the conditions failed*; trust the simulation there.

5. **Interval ⇔ test.** A 95% confidence interval for $\mu_1 - \mu_2$ is $(0.8, 3.4)$. Two-sided at $\alpha = 0.05$, does the matched t-test reject $H_0 : \mu_1 - \mu_2 = 0$?

   Show answer

   Yes. The interval lies entirely above 0, so 0 is not plausible — the matched two-sided test rejects H_0 at $\alpha = 0.05$.

Tech Tutorial: Compute

Tech Tutorial — Compute. This unit uses **formulas and theoretical distributions** (the t and the normal) instead of resampling. Pick **one** tool and follow its tab — all three handle this unit well.

We reuse `class_survey.csv`.

What you'll do

1. A **two-sample *t*-test** — do On- and Off-campus students differ in mean `sleep_hours`?
2. A **confidence interval for one proportion** — what fraction of students live on campus?

---

1. Two-sample *t*-test for a difference in means

This is the formula-based counterpart to the randomization test from the Simulate unit. It assumes roughly normal data (or a large enough sample) and reports a t statistic, degrees of freedom, and a p-value.

14.9.1 StatLens

1. Open the Two-Means Inference tool.
2. Confirm the **grouping** variable is `housing` and the **response** is `sleep_hours`.
3. Read the **t statistic, df, p-value**, and the confidence interval for the difference.

14.9.2 Jamovi

1. Open `class_survey.csv`.
2. **T-Tests → Independent Samples T-Test.** Dependent: `sleep_hours`; Grouping: `housing`.
3. Tick **Welch's** (unequal variances), **Mean difference**, and **Confidence interval**.

(*Jamovi screenshots to be added.*)

14.9.3 R

```
library(readr)
library(dplyr)
library(infer)

survey <- read_csv("../datasets/class_survey.csv")
```

```
# Two-sample t-test (Welch); order sets the sign of the difference
survey |>
  t_test(sleep_hours ~ housing,
         order = c("On-campus", "Off-campus"))
```

```
# A tibble: 1 x 7
  statistic  t_df p_value alternative estimate lower_ci upper_ci
      <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
1     0.204  47.8   0.839 two.sided     0.0516   -0.458    0.561
```

---

2. Confidence interval for one proportion

Here the variable is categorical: each student lives on campus or not. We estimate the population proportion living on campus with a normal-based (Wald/score) confidence interval.

14.9.4 StatLens

1. Open the One-Proportion Inference tool.
2. Choose `housing` and set the **success** category to `On-campus`.
3. Read the sample proportion and its **95% confidence interval**.

14.9.5 Jamovi

1. **Frequencies → 2 Outcomes (Binomial test).** Variable: `housing`.
2. Set the **test value** and read the **proportion** and its **confidence interval** (Jamovi reports the CI for the level you select).

(*Jamovi screenshots to be added.*)

14.9.6 R

```
# Count on-campus students and form a one-proportion z interval
on <- sum(survey$housing == "On-campus")
n  <- nrow(survey)

prop.test(on, n)$conf.int        # 95% CI for the proportion on campus
```

```
[1] 0.3026605 0.5865007
attr(,"conf.level")
[1] 0.95
```

```
round(on / n, 3)                 # sample proportion
```

```
[1] 0.44
```

Check yourself

- ☐ run a two-sample *t*-test and report t, df, p-value, and the CI for the difference;
- ☐ build a confidence interval for a single proportion;

☐ state, for each, the condition that makes the theoretical model appropriate.

Going further in R (optional)

OpenIntro *IMS* interactive tutorials for this unit (same `infer` style):

- Inference for a single proportion
- Inference for comparing two proportions
- Inference for difference in two means
- Introducing the t-distribution

Project: Inference for Proportions

Project — Inference for Proportions. A *project* is a small statistical analysis you carry out end to end and **communicate** as a short report. Do the analysis in **Jamovi** (or **R** — pick the tab that matches your section) and submit the deliverable described at the end.

The question

A 2011 trial tested whether the experimental **PfSPZ malaria vaccine** lowers the rate of infection. Twenty healthy volunteers were **randomly assigned** to receive either the vaccine or a placebo, then all were exposed to malaria under controlled conditions. Researchers recorded whether each volunteer became infected.

> **Does the vaccine reduce the proportion of people who become infected?**

The data

The `malaria` dataset (20 rows) has two categorical variables: `treatment` (`vaccine` or `placebo`) and `outcome` (`infection` or `no infection`). It is the kind of small randomized experiment where the **method of data collection** — random assignment — is what licenses a cause-and-effect conclusion.

Plan it first

Before you touch a tool, write down the plan. This is the part a calculator can't do for you.

a. **Parameter.** Define $p_1 - p_2$ in words (vaccine infection rate minus placebo infection rate).
b. **Hypotheses.** State H_0 and H_A for "the vaccine lowers the infection rate." One- or two-sided?
c. **Procedure.** This compares two groups on a yes/no outcome → a **two-proportion** comparison.
d. **Conditions.** Build the 2×2 table and check the success–failure counts. You will find the placebo group has **zero** non-infected people — so the normal-approximation z-procedure is **not** trustworthy here. That is your cue to rely on a **simulation-based (randomization) test**, the approach this course leads with.

> **Tip**
>
> This is a feature, not a snag: deciding *which* procedure the data can support is exactly the judgment the project is assessing. The honest answer here is "conditions for the z-test fail, so I'll use a randomization test."

Do the analysis

14.9.7 Jamovi

1. Open `malaria.csv` in Jamovi (**Open → Data**).
2. **Exploration → Contingency Tables → Independent Samples**. Put `treatment` in **Rows**, `outcome` in **Columns**. Read off the 2×2 counts and the row percentages (the infection rate in each group).
3. Under **Statistics**, the χ^2 test is reported; because a cell count is very small, also request **Fisher's exact test** — the small-sample-safe version. Record its p-value.
4. **Exploration → Descriptives**, split by `treatment`, or build a **bar plot** of `outcome` filled by `treatment` to show the two infection rates.

(*Jamovi screenshots to be added.*)

14.9.8 R

```
library(infer)
library(dplyr)
library(ggplot2)
library(openintro)
data(malaria)

# Infection rate in each group
malaria |>
  group_by(treatment) |>
  summarize(infected = mean(outcome == "infection"), n = n())
```

```
# A tibble: 2 x 3
  treatment infected     n
  <fct>        <dbl> <int>
1 placebo      1         6
2 vaccine      0.357    14
```

```
# Observed difference in infection proportions (vaccine - placebo)
obs_diff <- malaria |>
  specify(outcome ~ treatment, success = "infection") |>
  calculate(stat = "diff in props", order = c("vaccine", "placebo"))
obs_diff
```

```
Response: outcome (factor)
Explanatory: treatment (factor)
# A tibble: 1 x 1
    stat
   <dbl>
1 -0.643
```

```
# Randomization test: permute the treatment labels under "no effect"
set.seed(145)
null_dist <- malaria |>
  specify(outcome ~ treatment, success = "infection") |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "diff in props", order = c("vaccine", "placebo"))
```

```
# One-sided p-value: does the vaccine LOWER infection?
null_dist |> get_p_value(obs_stat = obs_diff, direction = "less")
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.015
```

```
# A graphic for the report
ggplot(malaria, aes(x = treatment, fill = outcome)) +
  geom_bar(position = "fill") +
  labs(x = "Treatment group", y = "Proportion", fill = "Outcome",
       title = "Infection rate by treatment group") +
  theme_minimal(base_size = 13)
```

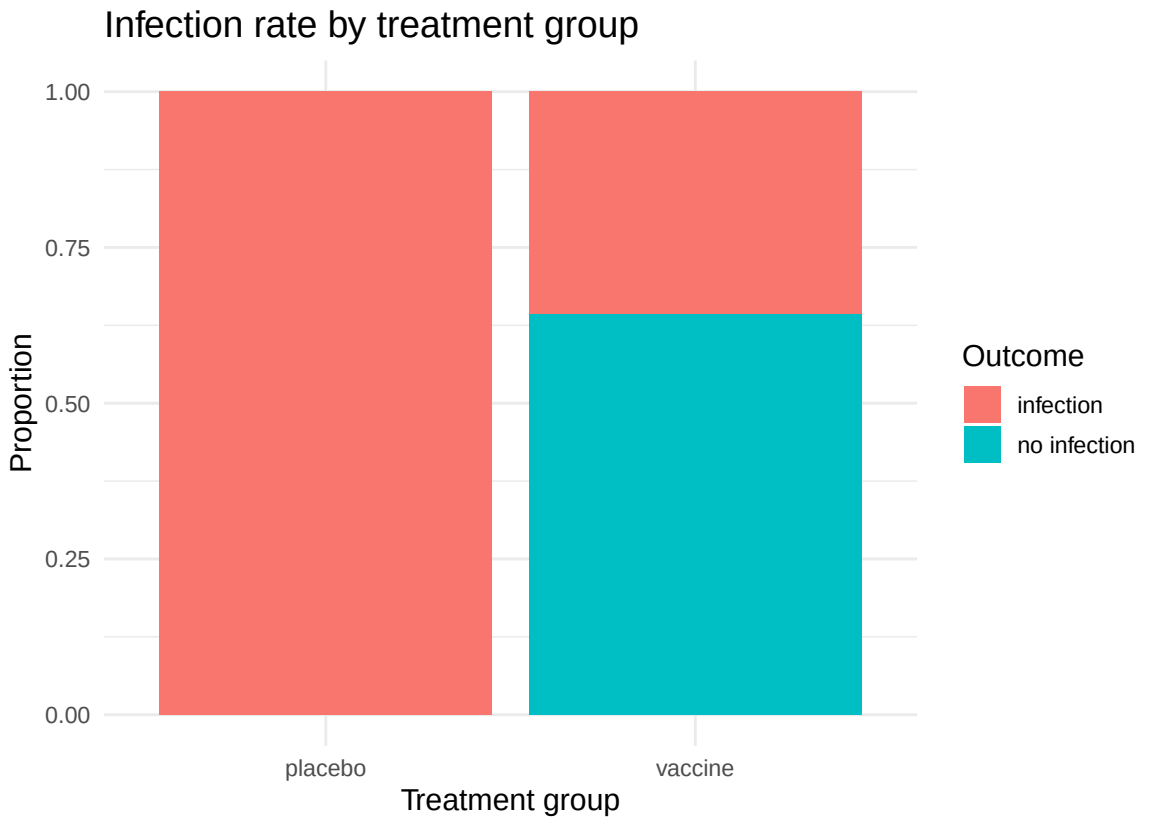


Figure 14.4: Infection rate by treatment group in the malaria vaccine trial. Every placebo volunteer became infected; the vaccine group's rate is much lower.

Your deliverable

Submit a short report containing **all** of:

a. **Graphic** — a labeled bar chart of the infection rate in each group.
b. **Numerical result** — the two infection proportions, their difference $\hat{p}_1 - \hat{p}_2$, and the **randomization p-value**.

c. **Conditions** — one sentence stating why the normal-approximation z-test is *not* appropriate here, and what you used instead.
d. **Conclusion in context** — 3–4 sentences answering the research question: the direction of the effect, whether the evidence is **discernible**, and — because volunteers were *randomly assigned* — whether a causal claim is justified.
e. **One limitation** — e.g., the very small sample, or generalizability beyond these 20 volunteers.

How it's graded

Criterion	What we look for
Appropriate, labeled graphic	infection rate by group, axes labeled
Procedure & conditions	recognizes the z-test fails; uses a randomization test
Correct numerical result	proportions, difference, and p-value reported correctly
Conclusion **in context**	answers the question, names direction + discernibility + causality
Communication	clear, correct interpretation; no probability-of-hypothesis fallacy

Want to see the idea first? Warm up with the malaria randomization activity before you build the deliverable — then come back and produce the report in Jamovi or R.

Project: Comparing Two Means

Project — Inference for Means. Carry out the analysis end to end in **Jamovi** (or **R**) and **communicate** the result as a short report. The graded skills are framing the comparison, checking conditions, and interpreting in context — not the arithmetic.

The question

Does a mother's smoking during pregnancy relate to her baby's birth weight?

> **Is the mean birth weight different for babies of smokers and nonsmokers?**

The data

The `ncbirths` dataset records 1,000 births in North Carolina. You'll use `weight` (birth weight, in pounds) and `habit` (`smoker` or `nonsmoker`). This is **observational** data — keep that in mind when you write the conclusion.

Plan it first

a. **Parameter.** Define $\mu_1 - \mu_2$ in words (mean weight for nonsmokers minus mean weight for smokers).
b. **Hypotheses.** State H_0 and H_A for "the mean birth weights differ." One- or two-sided?
c. **Procedure.** Two independent groups, a numerical outcome → a **two-sample** t procedure. (It is *not* paired — different babies in each group.)
d. **Conditions.** Each group is large and the data are not wildly skewed, so the t-procedure is reasonable. What would you check to be sure?

Do the analysis

14.9.9 Jamovi

1. Open `ncbirths.csv`.
2. **T-Tests → Independent Samples T-Test.** Dependent variable: `weight`; Grouping variable: `habit`.
3. Request **mean difference**, **confidence interval**, and a **descriptives** table; turn on the **normality** and **equality-of-variances** checks.
4. Read off the group means, the difference, the confidence interval, and the p-value.

(*Jamovi screenshots to be added.*)

14.9.10 R

```r
library(infer)
library(dplyr)
library(ggplot2)
library(openintro)
data(ncbirths)

# Keep rows with a recorded habit
d <- ncbirths |> filter(!is.na(habit))

# Group means, SDs, and sizes
d |>
  group_by(habit) |>
  summarize(mean = mean(weight), sd = sd(weight), n = n())
```

```
# A tibble: 2 x 4
  habit      mean    sd     n
  <fct>     <dbl> <dbl> <int>
1 nonsmoker  7.14  1.52   873
2 smoker     6.83  1.39   126
```

```r
# Two-sample t-test with a 95% CI for the difference (infer)
d |> t_test(weight ~ habit, order = c("nonsmoker", "smoker"))
```

```
# A tibble: 1 x 7
  statistic  t_df p_value alternative estimate lower_ci upper_ci
      <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
1      2.36  171.  0.0195 two.sided      0.316   0.0515    0.580
```

```r
ggplot(d, aes(habit, weight, fill = habit)) +
  geom_boxplot(show.legend = FALSE) +
  labs(x = "Maternal smoking", y = "Birth weight (lb)") +
  theme_minimal(base_size = 13)
```

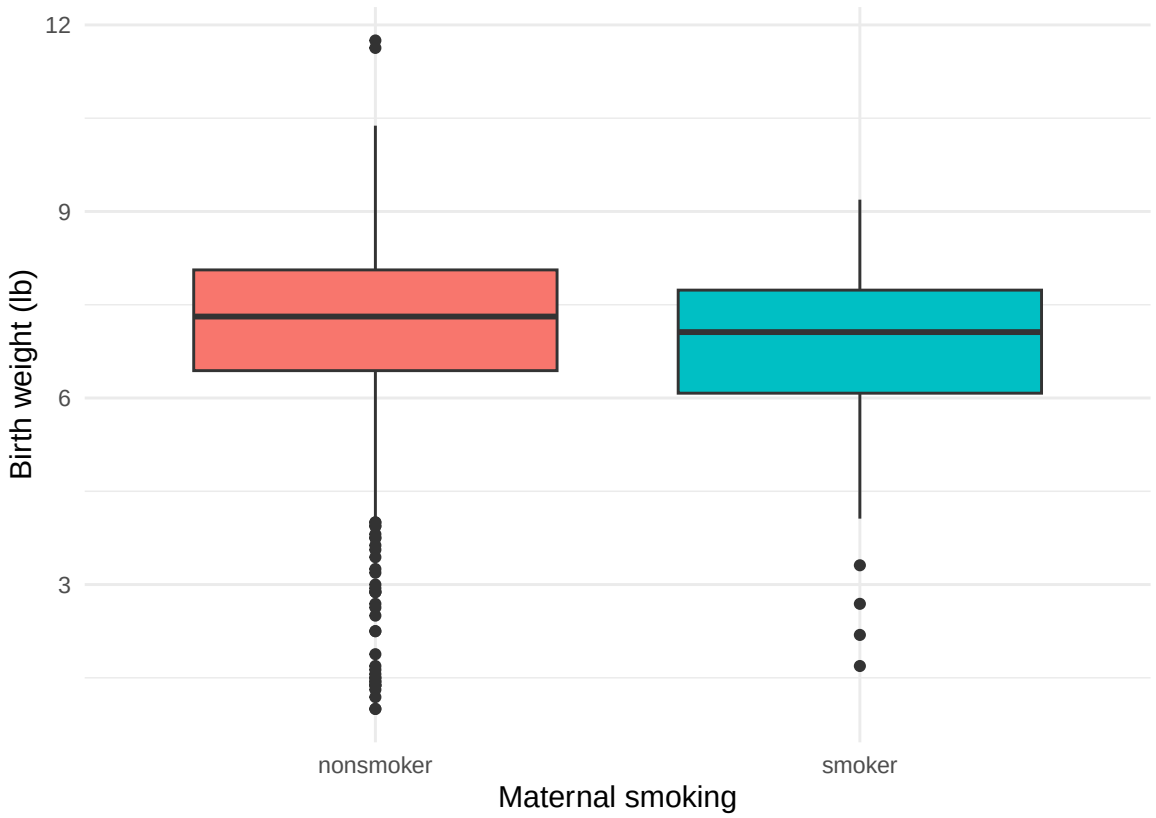


Figure 14.5: Birth weight by maternal smoking status. Babies of smokers weigh somewhat less on average.

Your deliverable

Submit a short report containing **all** of:

a. **Graphic** — labeled side-by-side boxplots of `weight` by `habit`.
b. **Numerical result** — the two group means, the difference, the 95% confidence interval, and the p-value.
c. **Conditions** — one sentence on why the two-sample t is reasonable here.
d. **Conclusion in context** — 3–4 sentences: is the difference **discernible**, in which direction, and — because this is **observational** — why you should *not* claim smoking *causes* the difference.
e. **One limitation** — e.g., confounding (smokers may differ in other ways), or self-reported smoking.

How it's graded

Criterion	What we look for
Appropriate, labeled graphic	boxplots by group, axes labeled
Procedure & conditions	two-sample t identified; conditions addressed
Correct numerical result	means, difference, CI, p-value reported correctly
Conclusion **in context**	discernibility + direction + **no causal claim** (observational)
Communication	clear, correct interpretation

Warm up first. Try the same comparison in the Two-Means tool before writing your report.

PART IV

UNIT 4: APPLY

You now have the full inference toolkit — simulation-based and formula-based. In this part, we apply that toolkit to new data types and richer questions: testing categorical distributions with chi-square, comparing multiple groups with ANOVA, and modeling relationships with correlation and linear regression.

Chapter 15

Chi-Square Goodness of Fit

In previous chapters, we have developed inferential methods for comparing proportions across two groups. But what if we want to assess whether a single categorical variable with multiple levels follows a particular distribution? For instance, we might want to know if a jury pool is racially representative of the community, or whether the days of the week on which accidents occur are equally likely. The **chi-square goodness of fit test** is designed for exactly this purpose. It compares a set of observed counts to a set of expected counts derived from a hypothesized distribution, using a single test statistic that summarizes all of the differences at once.

15.1 Observed versus expected counts

Consider data from a random sample of 275 jurors in a small county. Jurors identified their racial group, and we would like to determine if these jurors are racially representative of the county's population of registered voters. The observed counts and the population proportions are shown below.

	White	Black	Hispanic	Other	Total
Observed jurors	205	26	25	19	275
Population proportion	0.72	0.07	0.12	0.09	1.00

While the proportions in the jury sample do not precisely match the population proportions, it is unclear whether these data provide convincing evidence that the sample is not representative. If the jurors really were randomly sampled from the registered voters, we might expect small differences due to chance. However, unusually large differences may provide convincing evidence that the juries were not representative.

15.1.1 Computing expected counts

If the jurors were randomly sampled from the population (i.e., if the null hypothesis is true), we would expect the sample proportions to roughly match the population proportions. The **expected count** for each category is found by multiplying the sample size by the hypothesized proportion for that category.

Of the 275 people who served on a jury, about how many would we expect to be White if jurors are randomly selected from the population? How many would we expect to be Black?

About 72% of the population is White, so we would expect about 72% of the jurors to be White: $0.72 \times 275 = 198$. Similarly, we would expect about 7% of the jurors to be Black, which would correspond to about $0.07 \times 275 = 19.25$ Black jurors.

Twelve percent of the population is Hispanic and 9% represent other races. How many of the 275 jurors would we expect to be Hispanic? How many from other races?

Show answer

We would expect $0.12 \times 275 = 33$ Hispanic jurors and $0.09 \times 275 = 24.75$ jurors from other races.

The complete table of observed and expected counts is:

	White	Black	Hispanic	Other	Total
Observed counts	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

Computing expected counts for a goodness of fit test.

If the null hypothesis specifies proportions $p_1, p_2, \dots, p_k$ for k categories, and the total sample size is n, then the expected count for category i is:

$$E_i = n \times p_i$$

15.1.2 Setting up hypotheses

The sample proportions from each racial group were not a precise match for the population proportions. We need to test whether the differences are strong enough to provide convincing evidence that the jurors are not a random sample. These ideas can be organized into hypotheses:

- H_0: The jurors are a random sample from the population. The observed counts reflect natural sampling fluctuation.
- H_A: The jurors are not randomly sampled from the population. There is bias in juror selection.

To evaluate these hypotheses, we need to quantify how different the observed counts are from the expected counts. Strong evidence for the alternative hypothesis would come in the form of unusually large deviations from what would be expected based on sampling variation alone.

15.2 The chi-square statistic

To build a test statistic, we first compute the standardized difference between each observed and expected count. For each category i:

$$Z_i = \frac{O_i - E_i}{\sqrt{E_i}}$$

where O_i is the observed count and E_i is the expected count. The denominator $\sqrt{E_i}$ serves as the standard error for the count.

For the jury data:

$$Z_{\text{White}} = \frac{205 - 198}{\sqrt{198}} = 0.50$$
$$Z_{\text{Black}} = \frac{26 - 19.25}{\sqrt{19.25}} = 1.54$$
$$Z_{\text{Hispanic}} = \frac{25 - 33}{\sqrt{33}} = -1.39$$
$$Z_{\text{Other}} = \frac{19 - 24.75}{\sqrt{24.75}} = -1.16$$

We would like to use a single test statistic to determine if these four standardized differences are unusually far from zero as a group. We do this by squaring each standardized difference and adding them up:

$$X^2 = Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 0.50^2 + 1.54^2 + (-1.39)^2 + (-1.16)^2 = 5.89$$

Squaring each standardized difference before adding them together does two things:

1. Any standardized difference that is squared will be positive, so negative and positive deviations both contribute to the test statistic.
2. Differences that are already unusual (e.g., a standardized difference of 2.5) become much larger after squaring, which gives them more weight.

Chi-square test statistic.

The **chi-square test statistic** is the sum of the squared standardized differences between observed and expected counts:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where the sum is over all k categories, O_i is the observed count in category i, and E_i is the expected count in category i under the null hypothesis.

15.3 Simulation-based goodness of fit test

Before turning to the mathematical chi-square distribution, it is instructive to see how a simulation-based approach works for goodness of fit tests.

The idea is the same as in other randomization tests: if the null hypothesis is true (the data come from the hypothesized distribution), we can simulate many datasets that would arise under the null hypothesis and compute the chi-square statistic for each. This gives us a null distribution of the chi-square statistic, against which we can compare our observed statistic.

How the simulation works:

1. Under H_0, each observation has probability p_i of falling into category i.
2. For each simulated sample of size n, randomly assign each observation to a category according to the null probabilities.
3. Compute the chi-square statistic X^2 for the simulated data.
4. Repeat many times (e.g., 1,000 or 10,000 times) to build a null distribution.
5. Compare the observed X^2 to the null distribution. The p-value is the proportion of simulated statistics that are as large or larger than the observed statistic.

For the jury example, each simulated dataset would randomly assign 275 jurors to the four racial categories with probabilities 0.72, 0.07, 0.12, and 0.09. The resulting chi-square statistics from the simulated datasets form the null distribution.

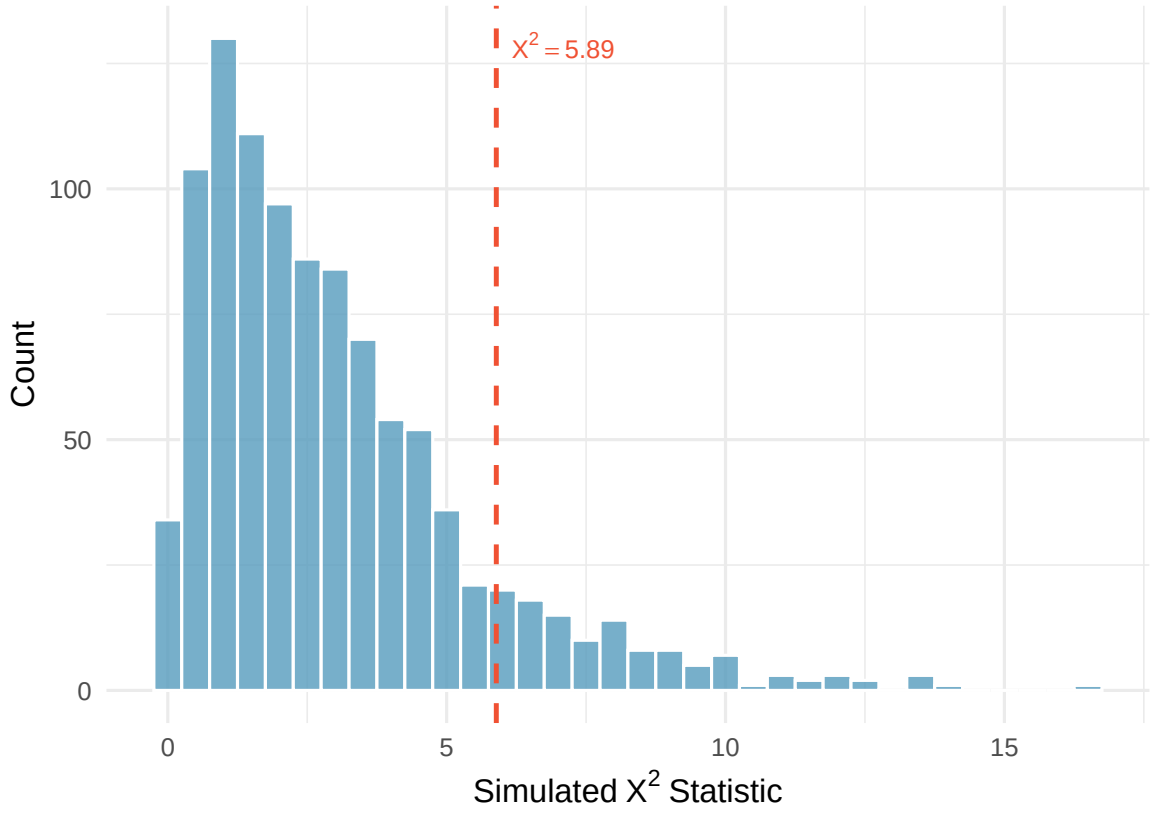


Figure 15.1: Histogram of chi-square statistics from 1,000 simulations under the null hypothesis for the jury example. The observed statistic of 5.89 is marked. Most simulated values fall below 5.89, but a meaningful fraction exceed it. Open in StatLens — the jury counts and null proportions are preloaded; regenerate this null distribution live.

The simulation-based approach confirms what we will find with the mathematical model: the observed chi-square statistic of 5.89 is not unusual enough to reject the null hypothesis at a 0.05 discernibility level.

See it in action. Open the Chi-Square GOF walkthrough for a guided pass through framing → checking expected counts → running the test → reading per-category contributions. Or open the Goodness-of-Fit Simulation to build the null χ^2 distribution by repeatedly drawing samples from the hypothesized distribution — a **Predict → Do → Explain** activity that lets you see *why* the mathematical chi-square curve is a reasonable model. For a bare tool, open the Goodness-of-Fit Test tool (analytic) or the Goodness-of-Fit Simulator (simulation) with the jury data preloaded, or enter your own data.

15.4 The chi-square distribution

The chi-square test statistic X^2 follows a **chi-square distribution** when the null hypothesis is true and the sample size conditions are met. The chi-square distribution is characterized by a single parameter called the **degrees of freedom** (df), which influences the shape, center, and spread of the distribution.

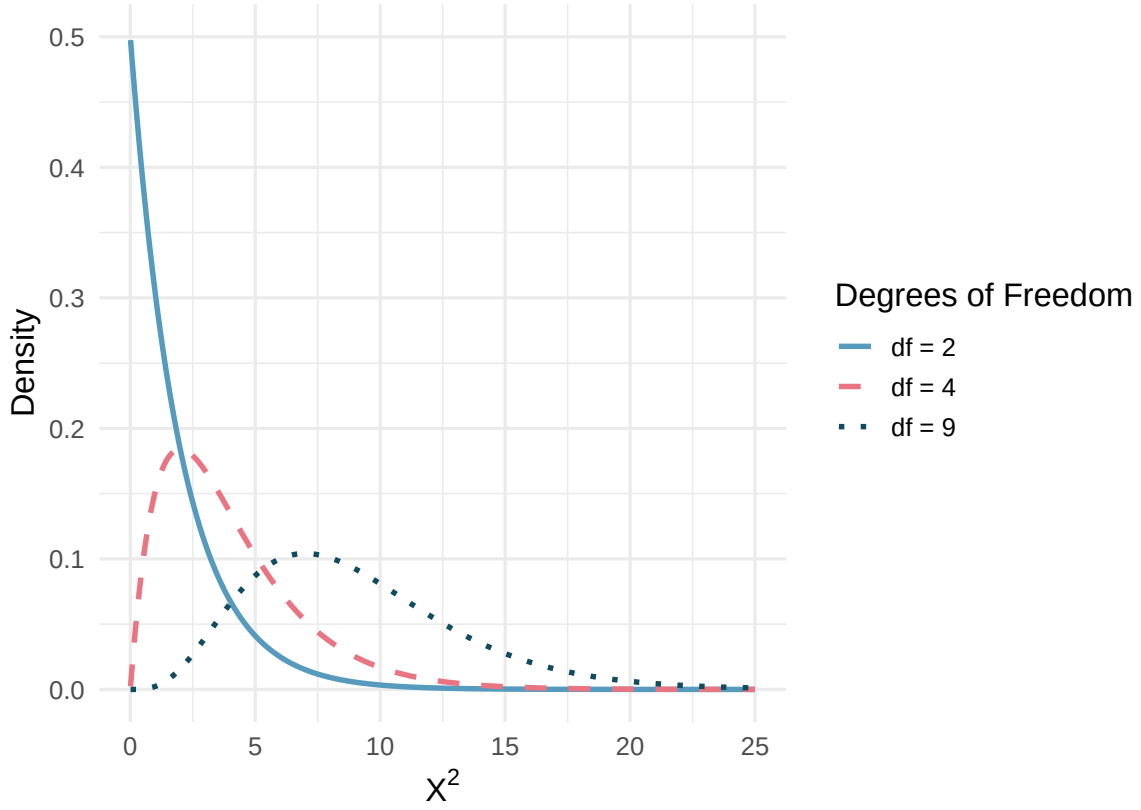


Figure 15.2: Three chi-square distributions with degrees of freedom 2, 4, and 9. As the degrees of freedom increase, the distribution becomes more symmetric, the center moves to the right, and the variability increases. ↗ Try this live — drive the df slider and watch the shape drift right.

Key properties of the chi-square distribution:

- It is always non-negative (the chi-square statistic is a sum of squared values).
- It is right-skewed, especially for small degrees of freedom.
- As the degrees of freedom increase, the distribution becomes more symmetric and shifts to the right.
- The mean of the distribution equals the degrees of freedom.

15.4.1 Degrees of freedom for goodness of fit

For a goodness of fit test with k categories, the degrees of freedom are:

$$df = k - 1$$

How many categories were there in the juror example? How many degrees of freedom should be associated with the chi-square distribution?

There were $k = 4$ categories: White, Black, Hispanic, and Other. The degrees of freedom are $df = k - 1 = 3$.

15.5 Finding a p-value

If the null hypothesis is true, X^2 follows a chi-square distribution with $k - 1$ degrees of freedom. Because larger values of X^2 correspond to greater deviations from the null hypothesis, the p-value is always found from the **upper tail** of the distribution.

For the jury data, the test statistic is $X^2 = 5.89$ and the degrees of freedom are $df = 3$. Find the p-value.

Using a chi-square distribution with 3 degrees of freedom, the area to the right of 5.89 is approximately 0.117.

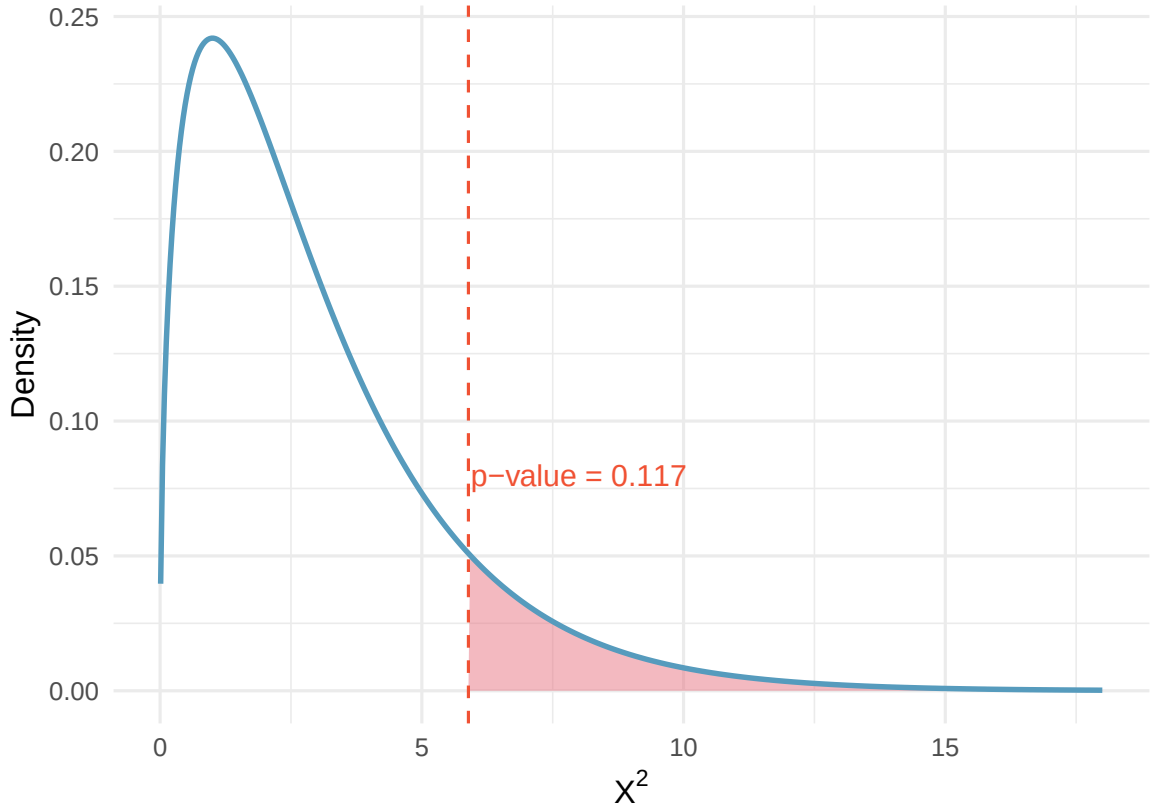


Figure 15.3: Chi-square distribution with 3 degrees of freedom, with the area above 5.89 shaded to represent the p-value of approximately 0.117. ↗ Try this live — drag the cutline to any χ^2 value and read the p-value.

Since the p-value of 0.117 is larger than $\alpha = 0.05$, we do not reject the null hypothesis. The data do not provide convincing evidence of racial bias in the juror selection process.

Consider a chi-square distribution with 5 degrees of freedom and a test statistic of 5.1. Find the p-value.

Using statistical software or a chi-square table, the area to the right of 5.1 on a chi-square distribution with 5 degrees of freedom is approximately 0.404. This is a very large p-value, so we would not reject the null hypothesis.

A chi-square distribution with 4 degrees of freedom has a test statistic of 10. Find the approximate p-value.

Show answer

Using statistical software, the area to the right of 10 on a chi-square distribution with 4 degrees of freedom is approximately 0.040. Since this is less than 0.05, we would reject the null hypothesis.

15.6 Chi-square goodness of fit test

We can now summarize the complete procedure for a chi-square goodness of fit test.

Chi-square test for goodness of fit.

Suppose we want to evaluate whether the observed counts $O_1, O_2, \dots, O_k$ in k categories are consistent with a hypothesized distribution that specifies expected counts $E_1, E_2, \dots, E_k$.

Hypotheses:

- H_0: The data follow the hypothesized distribution (the observed counts are consistent with the expected counts, up to sampling variability).
- H_A: The data do not follow the hypothesized distribution.

Test statistic:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

When the null hypothesis is true and the conditions are met, X^2 follows a chi-square distribution with $df = k - 1$.

The p-value is found from the upper tail of the chi-square distribution.

Conditions:

- **Independence.** Each case that contributes a count must be independent of all other cases.
- **Sample size.** Each expected count must be at least 5.

15.7 Evaluating goodness of fit for a distribution

The goodness of fit test can also be used to evaluate whether data follow a particular theoretical distribution.

Using 10 years of daily data for the S&P 500 stock index, we label each trading day as Up or Down and record how many days we must wait for each Up day: if the index rises on the next trading day, the wait is 1 day; if it falls once and then rises, the wait is 2 days; and so on. This gives 1,362 waiting times. We compare their observed frequencies to what we would expect if each day's movement were independent of previous days, with the same chance of an Up day each day (a geometric distribution model). The observed and expected counts are shown below.

Days until an Up day	1	2	3	4	5	6	7+
Observed	717	369	155	69	28	14	10
Expected	743	338	154	70	32	14	12

Are these data consistent with the geometric distribution model?

We set up the hypotheses:

- H_0: Stock returns are independent from one day to the next and follow a geometric model.
- H_A: Stock returns show a pattern inconsistent with the geometric model.

The chi-square statistic is:

$$X^2 = \frac{(717-743)^2}{743} + \frac{(369-338)^2}{338} + \frac{(155-154)^2}{154} + \frac{(69-70)^2}{70} + \frac{(28-32)^2}{32} + \frac{(14-14)^2}{14} + \frac{(10-12)^2}{12} = 4.61$$

With $k = 7$ categories, $df = 6$. The p-value for $X^2 = 4.61$ with 6 degrees of freedom is approximately 0.595.

Since the p-value is very large, we do not reject the null hypothesis. The data are consistent with the geometric distribution model, suggesting that stock movements may be independent from day to day over this period.

15.8 Chapter review

15.8.1 Summary

In this chapter, we developed the chi-square goodness of fit test, which assesses whether a sample of categorical data is consistent with a hypothesized distribution. The test uses the chi-square statistic, which measures how far the observed counts deviate from the expected counts. When the null hypothesis is true and the conditions are met, this statistic follows a chi-square distribution with $k-1$ degrees of freedom, where k is the number of categories. The p-value is always computed from the upper tail of the chi-square distribution, since larger values indicate greater discrepancy between the observed and expected counts.

15.8.2 Key terms

- Expected count
- Chi-square statistic (X^2)
- Chi-square distribution
- Degrees of freedom (goodness of fit)
- Goodness of fit test

15.9 Exercises

1. **True or false, Part I.** Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

a. The chi-square distribution, just like the normal distribution, has two parameters, mean and standard deviation.

b. The chi-square distribution is always right skewed, regardless of the value of the degrees of freedom parameter.

c. The chi-square statistic is always positive.

d. As the degrees of freedom increases, the shape of the chi-square distribution becomes more skewed.

2. **True or false, Part II.** Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

a. As the degrees of freedom increases, the mean of the chi-square distribution increases.

b. If you found $\chi^2 = 10$ with $df = 5$ you would fail to reject H_0 at the 5% significance level.

c. When finding the p-value of a chi-square test, we always shade the tail areas in both tails.

d. As the degrees of freedom increases, the variability of the chi-square distribution decreases.

3. **Open source textbook.** A professor using an open source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the semester he asks his students to complete a survey where they indicate what format of the book they used. Of the 126 students, 71 said they bought a hard copy of the book, 30 said they printed it out from the web, and 25 said they read it online.

a. State the hypotheses for testing if the professor's predictions were inaccurate.

b. How many students did the professor expect to buy the book, print the book, and read the book exclusively online?

c. This is an appropriate setting for a chi-square test. List the conditions required for a test and verify they are satisfied.

d. Calculate the chi-squared statistic, the degrees of freedom associated with it, and the p-value.

e. Based on the p-value calculated in part (d), what is the conclusion of the hypothesis test? Interpret your conclusion in this context.

4. **Barking deer.** Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined. In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7%, and deciduous forests make up 39.6%. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and 61 as deciduous forests. The table below summarizes these data.

Woods | Cultivated grassplot | Deciduous forests | Other | Total |

a. Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.

b. What type of test can we use to answer this research question?

c. Check if the assumptions and conditions required for this test are satisfied.

d. Do these data provide convincing evidence that barking deer prefer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Frame the question.** A candy company claims that its mixed-color bag contains 30% red, 20% orange, 20% yellow, 15% green, and 15% blue candies. A consumer group buys one large bag and counts $n = 400$ candies.
   a. What is the parameter framework here — a single proportion, several proportions, or a *distribution* across categories?
   b. Write the null and alternative hypotheses for the goodness-of-fit test in words *and* in symbols.
   c. Why is the GOF alternative always two-sided in spirit (there is no "one-sided" GOF)?
2. **Choose the procedure.** For each scenario, name the procedure — χ^2 **goodness-of-fit**, χ^2 **test of independence**, **one-proportion** z, or **not a chi-square problem**.
   a. Does the distribution of M&M colors in a bag match the manufacturer's claim?
   b. Is the gender distribution of voters (male/female/other) the same across three counties?
   c. Is the proportion of left-handed students at one school different from the national 10% rate?
   d. Are college majors (5 categories) independent of state of origin (4 categories)?
   e. Does birthday fall uniformly across the 12 months in a sample of 500 people?
3. **Check conditions before you trust the test.** Open the Goodness-of-Fit Test tool and enter (or load) a small dataset with at least one expected count below 5.
   a. What is the **expected-count condition** for the χ^2 GOF test? Why is it stated in terms of *expected* (not observed) counts?
   b. If one expected count is 3, should you trust the χ^2 p-value? What can you do instead — and what is the underlying issue with the χ^2 approximation in this regime?
   c. Suppose you have 10 categories and one expected count is 4. Can you simply merge the small-expected category with a neighbor? What is gained and what is lost by doing so?
4. **Run it and interpret it.** Open the Goodness-of-Fit Test tool and enter a five-category dataset (e.g., the candy bag).
   a. Report the χ^2 statistic, degrees of freedom, and p-value.
   b. State your decision at $\alpha = 0.05$ and write a one-sentence conclusion **in context**.
   c. The tool reports observed vs. expected counts for each category. Suppose H_0 is rejected — how do you tell *which* categories drove the rejection? What does a large positive "observed − expected" residual *for one category* tell you?
   d. Which statement is the correct interpretation of a p-value of 0.02?
      i. There is a 2% probability that the candy distribution matches the manufacturer's claim.
      ii. If the manufacturer's claim were true, we would see a fit at least this bad about 2% of the time by chance.
      iii. About 2% of candies in this bag came out wrong.
5. **Simulation as a sanity check.** Open the Goodness-of-Fit Simulator on the same dataset. (This tool draws repeated samples *from* the hypothesized distribution and builds the null χ^2 distribution empirically — the correct simulation analog for a GOF test.)
   a. Run 5000 simulated χ^2 statistics under H_0. What does the null distribution **look like** — shape, support, where the observed χ^2 falls?

b. Compare the **simulation** p-value to the χ^2 **approximation** p-value from Exercise 4. How close are they?
c. Now repeat on a small dataset where at least one expected count is below 5. Are the two p-values still close? Which one would you trust, and why?
d. A student says: "If the simulation and the formula agree, the simulation is redundant; if they disagree, the simulation is wrong." Why is this exactly backwards?

Chapter 16

Chi-Square Test of Independence

In the previous chapter, we used the chi-square statistic to test whether a single categorical variable follows a particular distribution. In this chapter, we extend the chi-square framework to two-way tables, where we have two categorical variables and want to assess whether they are associated. The key question becomes: is the distribution of one variable the same across all levels of the other variable, or does the distribution shift depending on the group? As with our other inference methods, we begin with a simulation-based approach (randomization) and then introduce the mathematical chi-square approximation.

Note that with two-way tables, there is not an obvious single parameter of interest. Instead, research questions focus on how the proportions of the response variable change (or not) across the different levels of the explanatory variable. Because there is no single population parameter to estimate, we focus on hypothesis testing (and not confidence intervals).

16.1 Randomization test of independence

16.1.1 Case study: The iPod experiment

We all buy used products – cars, computers, textbooks – and we sometimes assume the sellers will be forthright about any underlying problems. Researchers recruited 219 participants in a study where they would sell a used iPod that was known to have frozen twice in the past. The participants were incentivized to get as much money as they could. The researchers wanted to understand what types of questions would elicit the seller to disclose the freezing issue.

The buyers (who were collaborating with the researchers) asked one of three scripted questions:

- **General**: "What can you tell me about it?"
- **Positive Assumption**: "It does not have any problems, does it?"
- **Negative Assumption**: "What problems does it have?"

The results are shown below:

Question	Disclose problem	Hide problem	Total
General	2	71	73
Positive assumption	23	50	73
Negative assumption	36	37	73
Total	**61**	**158**	**219**

The data suggest that asking "What problems does it have?" was the most effective at getting the seller to disclose the past freezing issues. But could we see these results due to chance alone, or is this evidence that some questions are more effective at getting to the truth?

16.1.2 Expected counts in two-way tables

The hypothesis test is about assessing whether there is convincing evidence that the question asked was related to whether the seller disclosed the problem. In other words, we want to check whether the buyer's question was **independent** of whether the seller disclosed a problem.

From the experiment, the overall proportion of sellers who disclosed the freezing problem was $61/219 = 0.2785$. If there really is no difference among the questions and 27.85% of sellers were going to disclose regardless, how many of the 73 people in the General group would we have expected to disclose?

We would predict that $0.2785 \times 73 = 20.33$ sellers would disclose the problem. We observed far fewer than this, though it is not yet clear if that is due to chance variation or because the questions truly differ in effectiveness.

If the questions were actually equally effective, meaning about 27.85% of respondents would disclose the freezing issue regardless of what question they were asked, about how many sellers in the Positive Assumption group would we expect to *hide* the freezing problem?

Show answer

We would expect $(1 - 0.2785) \times 73 = 52.67$.

We can compute the expected number for all cells using the row totals, column totals, and table total. For instance, the proportion who disclosed overall was $61/219$. Multiplying by each row total:

$$\frac{61}{219} \times 73 = 20.33 \quad \text{(for each question group)}$$

More generally, this can be rewritten as:

$$\text{Expected Count}_{\text{row } i,\, \text{col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

Computing expected counts in a two-way table.

To calculate the expected count for the ith row and j^{th} column, compute:

$$\text{Expected Count}_{\text{row } i,\, \text{col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

Using this formula, the observed counts with expected counts in parentheses are:

Question	Disclose problem	Hide problem	Total
General	2 (20.33)	71 (52.67)	73
Positive assumption	23 (20.33)	50 (52.67)	73
Negative assumption	36 (20.33)	37 (52.67)	73
Total	**61**	**158**	**219**

16.1.3 The observed chi-square statistic

The chi-square test statistic for a two-way table is found by computing the squared standardized difference for every cell in the table and summing them:

$$\begin{array}{ll} \text{General formula} & \dfrac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \\ \text{Row 1, Col 1} & \dfrac{(2-20.33)^2}{20.33} = 16.53 \\ \text{Row 2, Col 1} & \dfrac{(23-20.33)^2}{20.33} = 0.35 \\ \vdots & \vdots \\ \text{Row 3, Col 2} & \dfrac{(37-52.67)^2}{52.67} = 4.66 \end{array}$$

Adding the computed value for each cell gives the chi-square test statistic:

$$X^2 = 16.53 + 0.35 + \cdots + 4.66 = 40.13$$

16.1.4 Variability of the statistic

Is 40.13 a big number? Does it indicate that the observed and expected values are really different, or is it a value we might expect to see just due to natural variability?

Assuming that the individuals would disclose or hide the problems **regardless** of the question they were given (i.e., that the null hypothesis is true), we can randomize the data by reassigning the 61 disclosed problems and 158 hidden problems to the three question groups at random. This is exactly the same type of randomization we used in previous chapters.

One such randomization might produce:

Question	Disclose problem	Hide problem	Total
General	29	44	73
Positive assumption	15	58	73
Negative assumption	17	56	73
Total	**61**	**158**	**219**

Computing the chi-square statistic for this randomized table:

$$X^2 = \frac{(29-20.33)^2}{20.33} + \frac{(15-20.33)^2}{20.33} + \cdots + \frac{(56-52.67)^2}{52.67} = 8$$

This is much smaller than 40.13.

16.1.5 Observed statistic vs. null chi-square statistics

One randomization is not sufficient. We generate many chi-square statistics under the null hypothesis to build a null distribution.

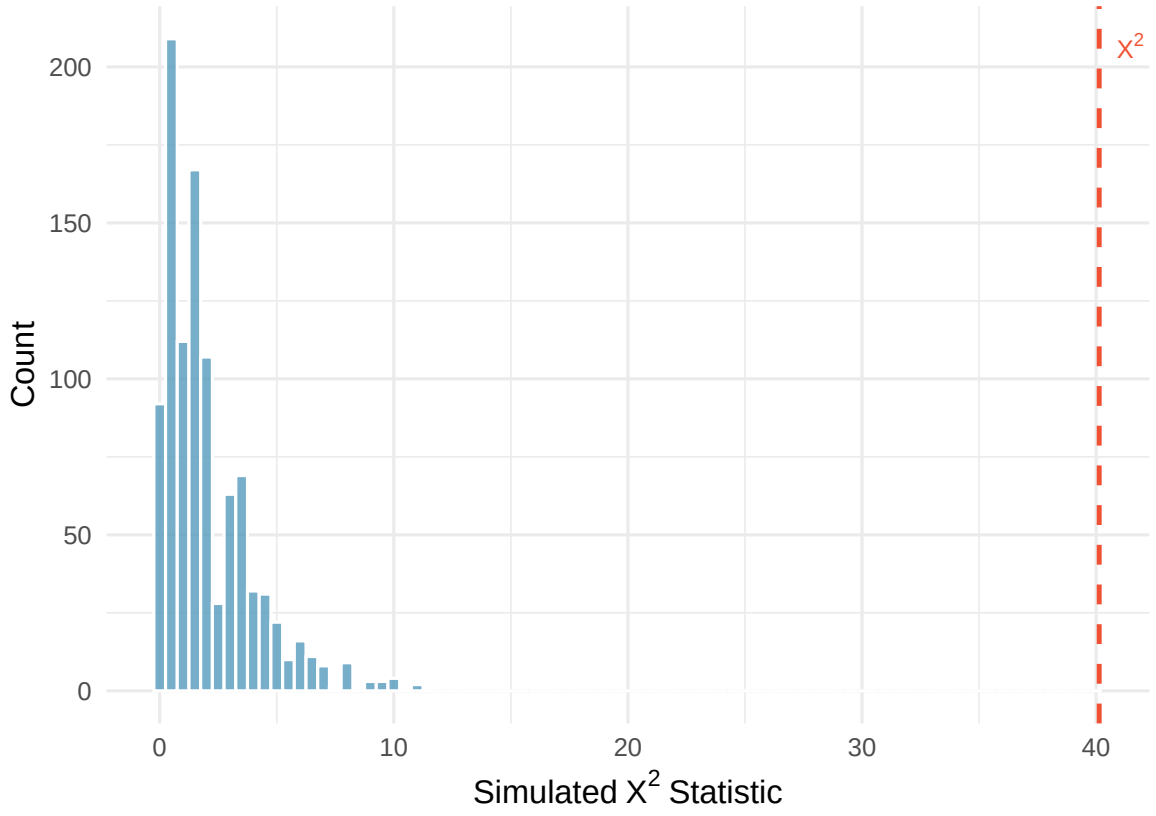


Figure 16.1: A histogram of chi-square statistics from 1,000 simulations produced under the null hypothesis, where the question is independent of the response. The observed statistic of 40.13 is marked by a red line. None of the 1,000 simulations had a chi-square value anywhere close to 40.13. ↗ Try this live — build the null χ^2 distribution live and place your observed value on it.

The observed value is so far from the null statistics that the simulated p-value is essentially zero. We conclude that the decision of whether to disclose the iPod's problem **is changed** by the question asked. We use the causal language of "changed" because the study was a randomized experiment.

Note that with a chi-square test, we only know that the two variables (question type and response) are related (i.e., not independent). We are not able to claim which specific type of question causes which specific type of response.

16.2 Mathematical model for test of independence

16.2.1 The chi-square test of independence

The chi-square test statistic follows a **chi-square distribution** when the null hypothesis is true and the conditions are met. For two-way tables, the degrees of freedom are:

$$df = (R-1) \times (C-1)$$

where R is the number of rows and C is the number of columns.

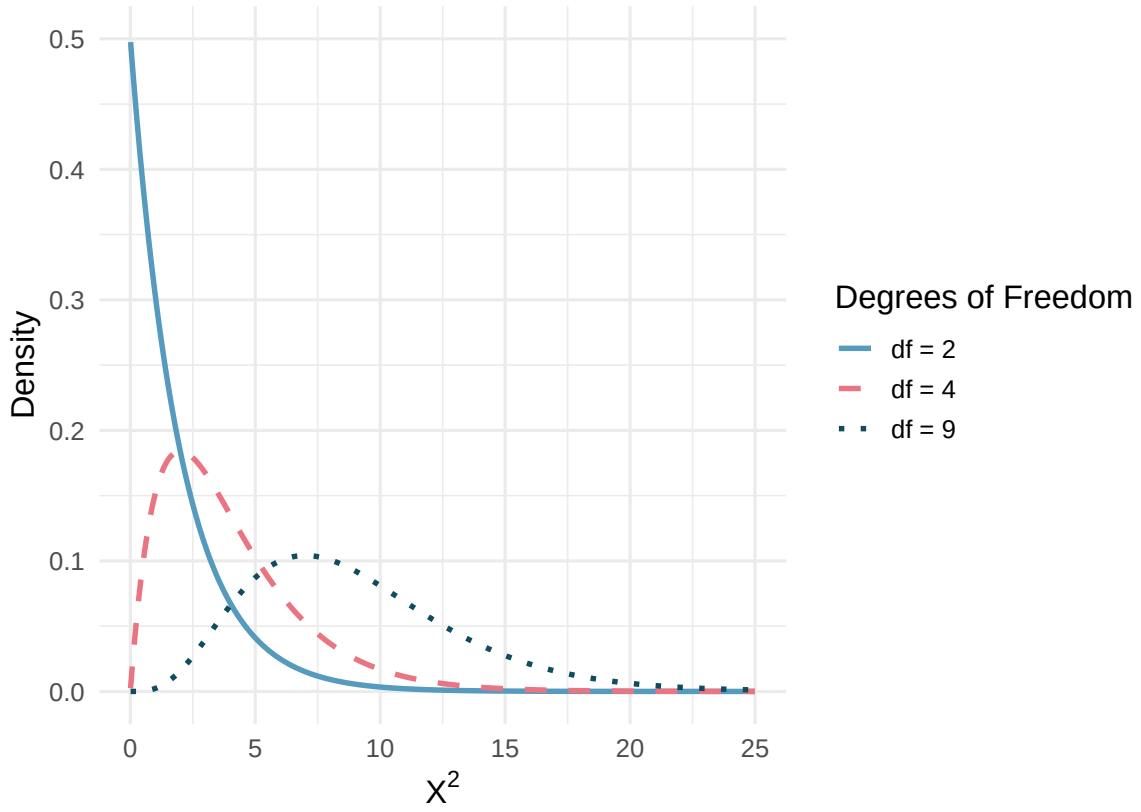


Figure 16.2: The chi-square distribution for different degrees of freedom. The larger the degrees of freedom, the longer the right tail extends. The smaller the degrees of freedom, the more peaked the mode on the left becomes. ↗ Try this live — drive the df slider and see the shape shift with df.

16.2.2 Variability of the chi-square statistic

As it turns out, the chi-square test statistic follows a chi-square distribution when the null hypothesis is true. For two-way tables, the degrees of freedom equal $(R-1) \times (C-1)$. In the iPod example, this is $(3-1) \times (2-1) = 2$.

The test statistic for assessing independence between two categorical variables is X^2.

The X^2 statistic is a ratio of how the observed counts differ from the expected counts, standardized by the expected counts:

$$X^2 = \sum_{i,j} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

When the null hypothesis is true and the conditions are met, X^2 has a chi-square distribution with $df = (R-1) \times (C-1)$.

Conditions:

- Independent observations
- Large samples: at least 5 expected counts in each cell

Computing degrees of freedom for a two-way table.

When applying the chi-square test to a two-way table, use $df = (R-1) \times (C-1)$ where R is the number of rows and C is the number of columns.

16.2.3 Finding the p-value

We can safely assume that the observations in the iPod experiment are independent (the question groups were randomly assigned). Each expected count exceeds 5, so the conditions for using the chi-square distribution are met. The test statistic $X^2 = 40.13$ follows a chi-square distribution with 2 degrees of freedom.

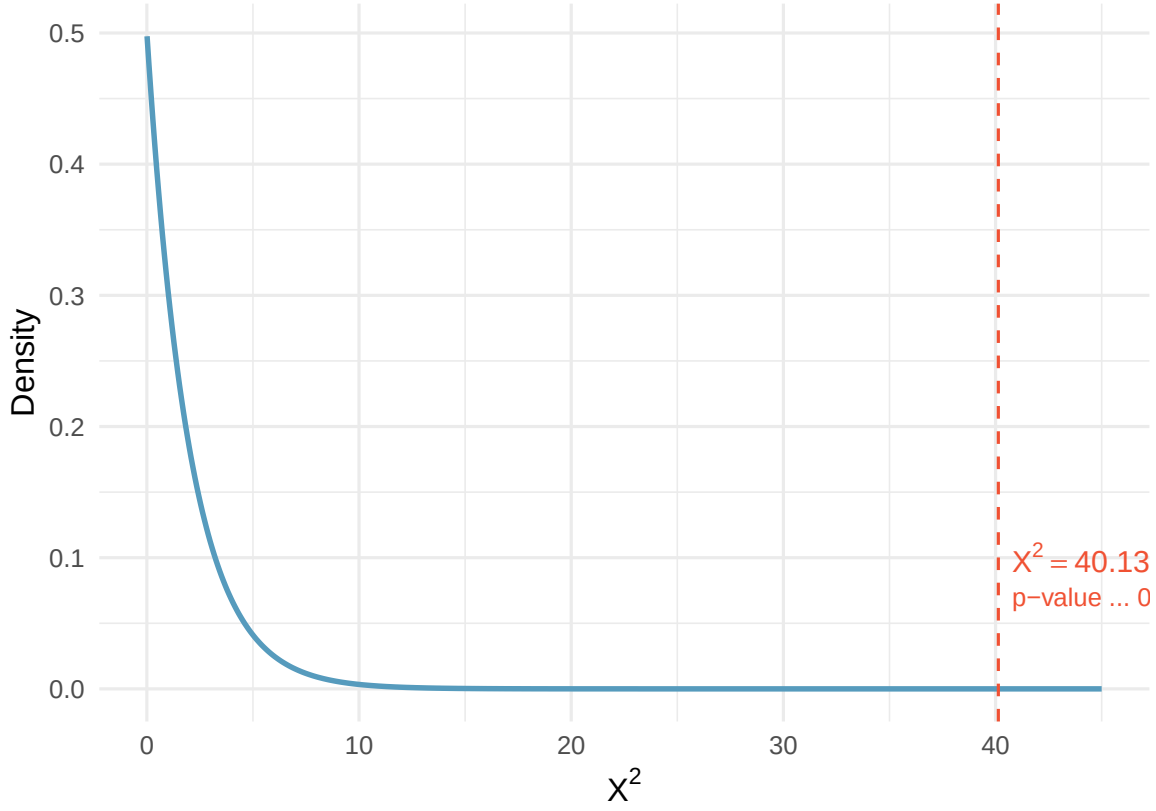


Figure 16.3: The chi-square distribution with df = 2, showing the area to the right of $\chi^2 = 40.13$. The p-value is so small it is not visible on the plot. ↗ Try this live — drag the cutline to any χ^2 value and read the p-value.

Find the p-value and draw a conclusion about whether the question affects the seller's likelihood of reporting the freezing problem.

Using a chi-square distribution with 2 degrees of freedom, the tail area above $X^2 = 40.13$ is approximately 0.000000002.

Using a discernibility level of $\alpha = 0.05$, the null hypothesis is rejected since the p-value is much smaller. The data provide convincing evidence that the question asked did affect a seller's likelihood to tell the truth about problems with the iPod.

16.2.4 Another example: Diabetes treatment

An experiment evaluated three treatments for Type 2 Diabetes in patients aged 10–17 who were being treated with metformin. The three treatments were: continued metformin alone (met), metformin combined with rosiglitazone (rosi), or a lifestyle intervention program. Each patient had a primary outcome of either failure (lacked glycemic control) or success (maintained glycemic control). The results are shown below.

Treatment	Failure	Success	Total
met	120	114	234
rosi	90	143	233
lifestyle	109	123	232
Total	**319**	**380**	**699**

What are appropriate hypotheses for this test?

- H_0: There is no difference in the effectiveness of the three treatments.
- H_A: There is some difference in effectiveness between the three treatments, e.g., perhaps the rosiglitazone treatment performed better than lifestyle intervention.

Compute the expected counts for each of the six cells in the diabetes treatment table.

Show answer

Row 1, Col 1: $\frac{234 \times 319}{699} = 106.8$. Row 1, Col 2: $\frac{234 \times 380}{699} = 127.2$. Row 2, Col 1: $\frac{233 \times 319}{699} = 106.3$. Row 2, Col 2: $\frac{233 \times 380}{699} = 126.7$. Row 3, Col 1: $\frac{232 \times 319}{699} = 105.9$. Row 3, Col 2: $\frac{232 \times 380}{699} = 126.1$.

Note: when analyzing 2-by-2 contingency tables (where both variables have only two categories), it is generally preferable to use the two-proportion methods from earlier chapters.

See it in action. Open the Chi-Square Independence walkthrough for a guided pass through framing → conditions → running the test → reading standardized residuals. Or open the Chi-Square Test or Randomization Chi-Square directly on your own data.

16.3 Comparing the simulation and chi-square approaches

Both the randomization test and the mathematical chi-square test address the same question: are two categorical variables independent? The two approaches generally give very similar results, especially when the sample size is large and the expected counts are well above 5.

The **randomization test** is more flexible because:

- It does not require any minimum expected count.
- It works well even for small samples.
- It requires no distributional assumptions beyond independence.

The **mathematical chi-square test** is more convenient because:

- It does not require a computer to simulate thousands of randomizations.
- It provides a simple formula and distribution for computing p-values.
- It is widely implemented in statistical software.

In practice, the chi-square test is the standard approach when conditions are met, while the randomization test provides a useful check or alternative when conditions are questionable.

16.4 Chapter review

16.4.1 Summary

In this chapter, we extended the chi-square framework to two-way tables, testing whether two categorical variables are independent. The randomization approach shuffles the data under the null hypothesis to build a distribution of the chi-square statistic, while the mathematical approach uses the chi-square distribution with $(R-1) \times (C-1)$ degrees of freedom. Both approaches require that observations are independent and that expected counts are sufficiently large (at least 5 in each cell).

The chi-square test of independence tells us *whether* two variables are related, but does not tell us *how* they are related or *which* groups differ. To understand the nature of the relationship, we must examine the individual cells and their contributions to the chi-square statistic.

16.4.2 Key terms

- Independence
- Expected counts (two-way table)
- Chi-square statistic
- Chi-square distribution
- Degrees of freedom (two-way table)

16.5 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Quitters.** Does being part of a support group affect the ability of people to quit smoking? A county health department enrolled 300 smokers in a randomized experiment. 150 participants were randomly assigned to a group that used a nicotine patch and met weekly with a support group; the other 150 received the patch and did not meet with a support group. At the end of the study, 40 of the participants in the patch plus support group had quit smoking while only 30 smokers had quit in the other group.

a. Create a two-way table presenting the results of this study.

b. Answer each of the following questions under the null hypothesis that being part of a support group does not affect the ability of people to quit smoking, and indicate whether the expected values are higher or lower than the observed values.

   i. How many subjects in the "patch + support" group would you expect to quit?

   ii. How many subjects in the "patch only" group would you expect to not quit?

2. **Act on climate change.** The table below summarizes results from a Pew Research poll which asked respondents whether they have personally taken action to help address climate change within the last year and their generation. The differences in each generational group may be due to chance. Complete the following computations under the null hypothesis of independence between an individual's generation and whether they have personally taken action to help address climate change within the last year. (Pew Research Center 2021a)

	Response		
Generation	Took action	Didn't take action	Total
Gen Z	292	620	912
Millenial	885	2,275	3,160
Gen X	809	2,709	3,518
Boomer & older	1,276	4,798	6,074
Total	3,262	10,402	13,664

If there is no relationship between age and action,

a. how many Gen Z'ers would you expect to have personally taken action to help address climate change within the last year?

b. how many Millenials would you expect to have personally taken action to help address climate change within the last year?

c. how many Gen X'ers would you expect to have personally taken action to help address climate change within the last year?

d. how many Boomers and older would you expect to have personally taken action to help address climate change within the last year?

3. **Lizard habitats, data.** In order to assess whether habitat conditions are related to the sunlight choices a lizard makes for resting, Western fence lizard (*Sceloporus occidentalis*) were observed across three different microhabitats. (Adolph 1990; Asbury and Adolph 2007)

	sunlight			
site	sun	partial	shade	Total
desert	16	32	71	119
mountain	56	36	15	107
valley	42	40	24	106
Total	114	108	110	332

   a. If the variables describing the habitat and the amount of sunlight are independent, what proporiton of lizards (total) would be expected in each of the three sunlight categories?

   b. Given the proportions of each sunlight condition, how many lizards of each type would you expect to see in the sun? in the partial sun? in the shade?

   c. Compare the observed (original data) and expected (part b.) tables. From a first glance, does it seem as though the habitat and choice of sunlight may be associated?

   d. Regardless of your answer to part (c), is it possible to tell from looking only at the expected and observed counts whether the two variables are associated?

4. **Disaggregating Asian American tobacco use, data.** Understanding cultural differences in tobacco use across different demographic groups can lead to improved health care education and treatment. A recent study disaggregated tobacco use across Asian American ethnic groups including Asian-Indian (n = 4,373), Chinese (n = 4,736), and Filipino (n = 4,912), in comparison to non-Hispanic Whites (n = 275,025). The number of current smokers in each group was reported as Asian-Indian (n = 223), Chinese (n = 279), Filipino (n = 609), and non-Hispanic Whites (n = 50,880). (Rao et al. 2021)

In order to assess whether there is a difference in current smoking rates across three Asian American ethnic groups, the observed data is compared to the data that would be expected if there were no association between the variables.

	Smoking		
ethnicity	don't smoke	smoke	Total
Asian-Indian	4,150	223	4,373
Chinese	4,457	279	4,736
Filipino	4,303	609	4,912
Total	12,910	1,111	14,021

   a. If the variables on ethnicity and smoking status are independent, estimate the proporiton of individuals (total) who smoke?

   b. Given the overall proportion who smoke, how many of each Asian American ethnicity would you expect to smoke?

   c. Compare the observed and expected counts. From a first glance, does it seem as though the Asian American ethnicity and choice of smoking may be associated?

   d. Regardless of your answer to part (c), is it possible to tell from looking only at the expected and observed counts whether the two variables are associated?

5. **Lizard habitats, randomize once.** In order to assess whether habitat conditions are related to the sunlight choices a lizard makes for resting, Western fence lizard (*Sceloporus occidentalis*) were observed across three different microhabitats. (Adolph 1990; Asbury and Adolph 2007) Then, the data were randomized once, where sunlight preference was randomly assigned to the lizards across different sites. The original data are shown on the left and the results of the randomization is shown on the right.

Original data				
	sunlight			
site	sun	partial	shade	Total
desert	16	32	71	119
mountain	56	36	15	107
valley	42	40	24	106
Total	114	108	110	332

Randomized data				
	sunlight			
site	sun	partial	shade	Total
desert	44	42	33	119
mountain	39	31	37	107
valley	31	35	40	106
Total	114	108	110	332

Recall that the Chi-squared statistic (X^2) measures the difference between the expected and observed counts. Without calculating the actual statistic, report on whether the original data or the randomized data will have a larger Chi-squared statistic. Explain your choice.

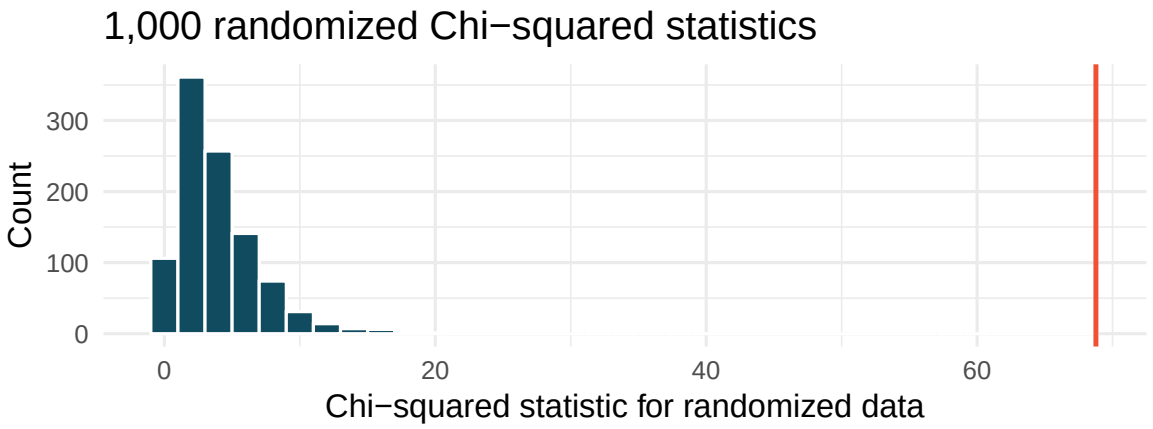
6. **Disaggregating Asian American tobacco use, randomize once.** In a study that aims to disaggregate tobacco use across Asian American ethnic groups (Asian-Indian, Chinese, and Filipino, in comparison to non-Hispanic Whites), respondents were asked whether they smoke tobacco or not. (Rao et al. 2021) Then, the data were randomized once, where smoking status was randomly assigned to the participants across different ethnicities. The original data are shown on the left and the results of the randomization is shown on the right.

Original data			
	Smoking		
ethnicity	don't smoke	smoke	Total
Asian-Indian	4,150	223	4,373
Chinese	4,457	279	4,736
Filipino	4,303	609	4,912
Total	12,910	1,111	14,021

Randomized data			
	Smoking		
ethnicity	don't smoke	smoke	Total
Asian-Indian	4,015	358	4,373
Chinese	4,385	351	4,736
Filipino	4,510	402	4,912
Total	12,910	1,111	14,021

Recall that the Chi-squared statistic (X^2) measures the difference between the expected and observed counts. Without calculating the actual statistic, report on whether the original data or the randomized data will have a larger Chi-squared statistic. Explain your choice.

7. **Lizard habitats, randomization test.** In order to assess whether habitat conditions are related to the sunlight choices a lizard makes for resting, Western fence lizard (*Sceloporus occidentalis*) were observed across three different microhabitats. (Adolph 1990; Asbury and Adolph 2007) The original data were randomized 1,000 times (sunlight variable randomly assigned to the observations across different habitats), and the histogram of the Chi-squared statistic on each randomization is displayed.

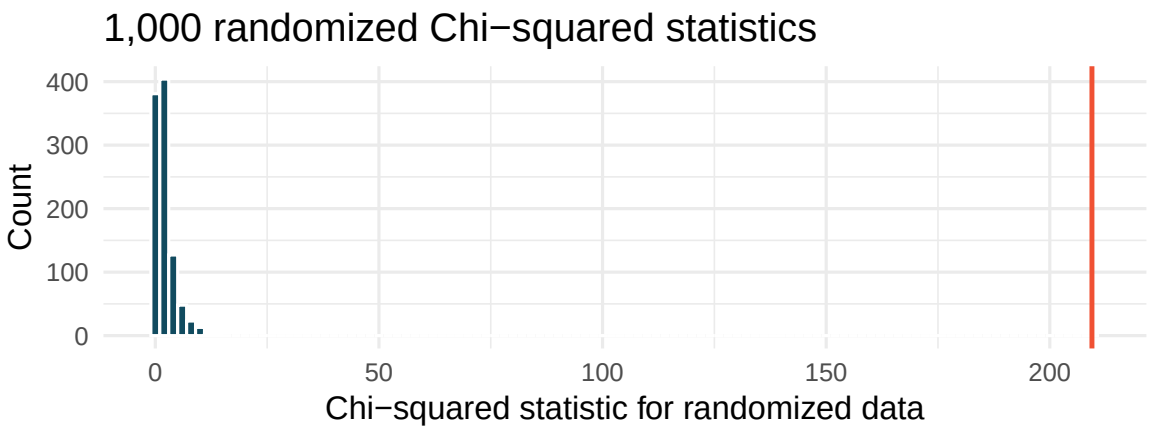


a. The histogram above describes the Chi-squared statistics for 1,000 different randomization datasets. When randomizing the data, is the imposed structure that the variables are independent or that the variables are associated? Explain.

b. What is the range of plausible values for the randomized Chi-squared statistic?

c. The observed Chi-squared statistic is 68.8 (marked in red on plot). Does the observed value provide evidence against the null hypothesis? To answer the question, state the null and alternative hypotheses, approximate the p-value, and conclude the test in the context of the problem.

8. **Disaggregating Asian American tobacco use, randomization test.** Understanding cultural differences in tobacco use across different demographic groups can lead to improved health care education and treatment. A recent study disaggregated tobacco use across Asian American ethnic groups including Asian-Indian (n = 4373), Chinese (n = 4736), and Filipino (n = 4912), in comparison to non-Hispanic Whites (n = 275,025). The number of current smokers in each group was reported as Asian-Indian (n = 223), Chinese (n = 279), Filipino (n = 609), and non-Hispanic Whites (n = 50,880). (Rao et al. 2021) The original data were randomized 1000 times (smoking status randomly assigned to the observations across ethnicities), and the histogram of the Chi-squared statistic on each randomization is displayed.



a. The histogram above describes the Chi-squared statistics for 1000 different randomization datasets. When randomizing the data, is the imposed structure that the variables are independent or that the variables are associated? Explain.

b. What is the range of plausible values for the randomized Chi-squared statistic?

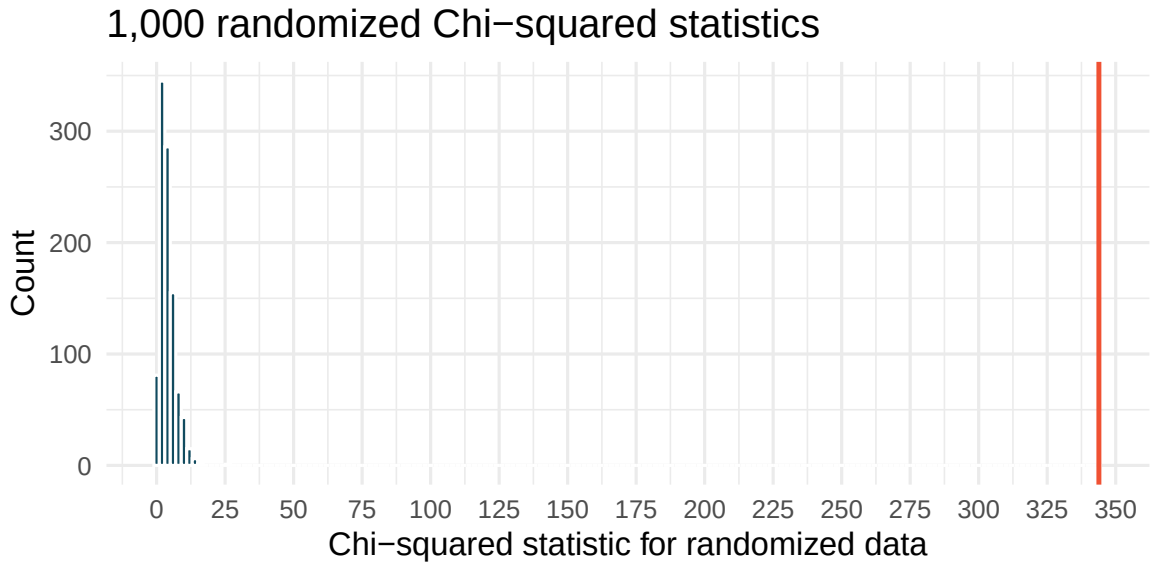
c. The observed Chi-squared statistic is 209.42 (marked in red on plot). Does the observed value provide evidence against the null hypothesis? To answer the question, state the null and alternative hypotheses, approximate the p-value, and conclude the test in the context of the problem.

9. **Lizard habitats, larger data.** In order to assess whether habitat conditions are related to the sunlight choices a lizard makes for resting, Western fence lizard (*Sceloporus occidentalis*) were observed across three different microhabitats. (Adolph 1990; Asbury and Adolph 2007)

Consider the situation where the dataset is 5 times *larger* than the original data (but have the same proportional representation in each category). The distribution of lizards in each of the sites resting in the sun, partial sun, and shade are as follows.

Larger data				
	sunlight			
site	sun	partial	shade	Total
desert	80	160	355	595
mountain	280	180	75	535
valley	210	200	120	530
Total	570	540	550	1,660

The larger dataset was randomized 1,000 times (sunlight preference randomly assigned to the observations across sites), and the histogram of the Chi-squared statistic on each randomization is displayed.



a. The histogram above describes the Chi-squared statistics for 1,000 different randomization of the larger dataset. When randomizing the data, is the imposed structure that the variables are independent or that the variables are associated? Explain.

b. What is the (approximate) range of plausible values for the randomized Chi-squared statistic?

c. The observed Chi-squared statistic is 343.865 (and seen in red on the graph). Does the observed value provide evidence against the null hypothesis? To answer the question, state the null and alternative hypotheses, approximate the p-value, and conclude the test in the context of the problem.

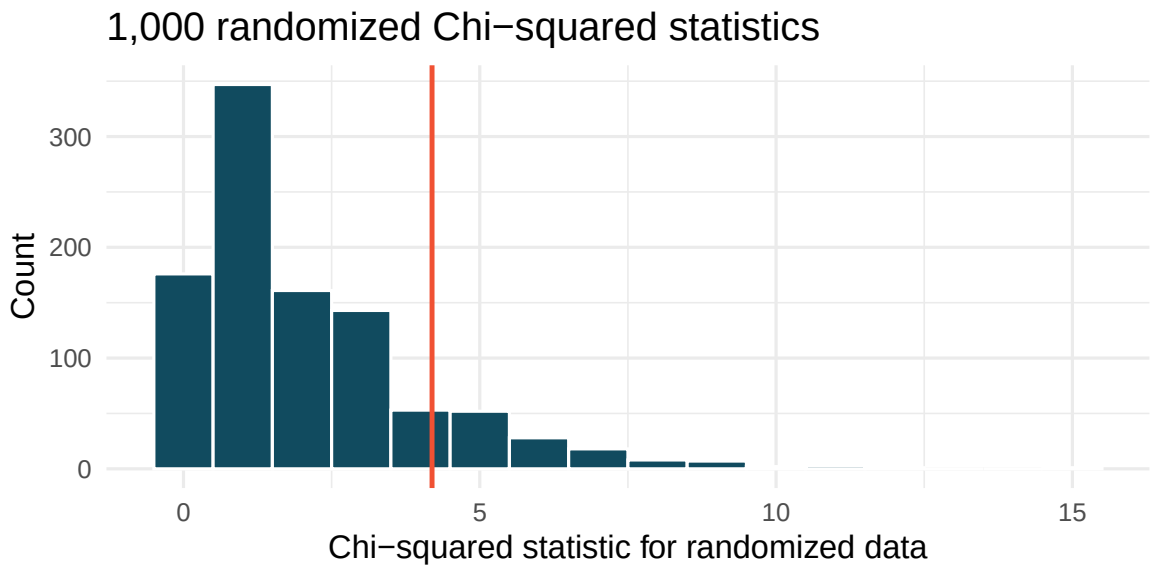
d. If the alternative hypothesis is true, how does the sample size effect the ability to reject the null hypothesis? (*Hint:* Consider the original data as compared with the larger dataset that have the same proportional values.)

10. **Disaggregating Asian American tobacco use, smaller data.** Understanding cultural differences in tobacco use across different demographic groups can lead to improved health care education and treatment. A recent study disaggregated tobacco use across Asian American ethnic groups. (Rao et al. 2021)

Consider the situation where the dataset is 50 times *smaller* than the original data (but have the same proportional representation in each category). The distribution of smokers in each of the ethnicity groups in the smaller data are as follows.

Smaller data			
	Smoking		
ethnicity	don't smoke	smoke	Total
Asian-Indian	83	4	87
Chinese	89	6	95
Filipino	86	12	98
Total	258	22	280

The smaller dataset was randomized 1,000 times (smoking status randomly assigned to the observations across ethnicities), and the histogram of the Chi-squared statistic on each randomization is displayed.



a. The histogram above describes the Chi-squared statistics for 1,000 different randomization of the smaller dataset. When randomizing the data, is the imposed structure that the variables are independent or that the variables are associated? Explain.

b. What is the (approximate) range of plausible values for the randomized Chi-squared statistic?

c. The observed Chi-squared statistic is 4.19 (and seen in red on the graph). Does the observed value provide evidence against the null hypothesis? To answer the question, state the null and alternative hypotheses, approximate the p-value, and conclude the test in the context of the problem.

d. If the alternative hypothesis is true, how does the sample size effect the ability to reject the null hypothesis? (*Hint:* Consider the original data as compared with the smaller dataset that have the same proportional values.)

11. **True / False, I.** Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

a. The Chi-square distribution, just like the normal distribution, has two parameters, mean and standard deviation.

b. The Chi-square distribution is always right skewed, regardless of the value of the degrees of freedom parameter.

c. The Chi-square statistic is always greater than or equal to 0.

d. As the degrees of freedom increases, the shape of the Chi-square distribution becomes more skewed.

12. **True / False, II.** Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

a. As the degrees of freedom increases, the mean of the Chi-square distribution increases.

b. If you found $\chi^2 = 10$ with $df = 5$ you would fail to reject H_0 at the 5% discernibility level.

c. When finding the p-value of a Chi-square test, we always shade the tail areas in both tails.

d. As the degrees of freedom increases, the variability of the Chi-square distribution decreases.

13. **Sleep deprived transportation workers.** The National Sleep Foundation conducted a survey on the sleep habits of randomly sampled transportation workers and randomly sampled non-transportation workers that serve as a "control" for comparison. (National Sleep Foundation 2012) The results of the survey are shown below. Conduct a hypothesis test to evaluate if these data provide evidence of an association between sleep levels and profession.

Profession	Less than 6 hours	6 to 8 hours	More than 8 hours	Total
Non-transportation workers	35	193	64	292
Transportation workers	104	499	192	795
Total	139	692	256	1,087

14. **Parasitic worm.** Lymphatic filariasis is a disease caused by a parasitic worm. Complications of the disease can lead to extreme swelling and other complications. Here we consider results from a randomized experiment that compared three different drug treatment options to clear people of the this parasite, which people are working to eliminate entirely. The results for the second year of the study are given below: (King et al. 2018)

	Outcome		
group	Clear at Year 2	Not Clear at Year 2	Total
Three drugs	52	2	54
Two drugs	31	24	55
Two drugs annually	42	14	56
Total	125	40	165

a. Set up hypotheses for evaluating whether there is any difference in the performance of the treatments, and also check conditions.

b. Statistical software was used to run a Chi-square test, which output: $X^2 = 23.7 \quad df = 2 \quad$ p-value < 0.0001. Use these results to evaluate the hypotheses from part (a), and provide a conclusion in the context of the problem.

15. **Shipping holiday gifts.** A local news survey asked 500 randomly sampled Los Angeles residents which shipping carrier they prefer to use for shipping holiday gifts. The table below shows the distribution of responses by age group as well as the expected counts for each cell (shown in italics).

Shipping method	Age						Total
	18-34		35-54		55+		
USPS	72	*81*	97	*102*	76	*62*	245
UPS	52	*53*	76	*68*	34	*41*	162
FedEx	31	*21*	24	*27*	9	*16*	64
Something else	7	*5*	6	*7*	3	*4*	16
Not sure	3	*5*	6	*5*	4	*3*	13
Total	165		209		126		500

a. State the null and alternative hypotheses for testing for independence of age and preferred shipping method for holiday gifts among Los Angeles residents.

b. Are the conditions for inference using a Chi-square test satisfied?

16. **Coffee and depression.** Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician- diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption. (Lucas et al. 2011)

Clinical depression	Caffeinated coffee consumption					Total
	1 cup / week or fewer	2-6 cups / week	1 cups / day	2-3 cups / day	4 cups / day or more	
Yes	670	___	905	564	95	2,607
No	11,545	6,244	16,329	11,726	2,288	48,132
Total	12,215	6,617	17,234	12,290	2,383	50,739

a. What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

b. Write the hypotheses for the test you identified in part (a).

c. Calculate the overall proportion of women who do and do not suffer from depression.

d. Identify the expected count for the empty cell, and calculate the contribution of this cell to the test statistic.

e. The test statistic is $\chi^2 = 20.93$. What is the p-value?

f. What is the conclusion of the hypothesis test?

g. One of the authors of this study was quoted on the New York Times as saying it was "too early to recommend that women load up on extra coffee" based on just this study. (O'Connor 2011) Do you agree with this statement? Explain your reasoning.

Datasets: `lizard_habitat` (openintro)

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Frame the question.** A campus survey crosses **major** (STEM / humanities / social science / business) with **caffeine source** (coffee / tea / energy drink / none) for 600 students.
   a. Identify the two categorical variables and the dimensions of the contingency table.
   b. Write the null and alternative hypotheses for a test of independence, in words.
   c. The test of **independence** and the test of **homogeneity** use the same χ^2 statistic, the same expected-count formula, and the same df. What is the *conceptual* difference between them, and why does it matter how the data were collected?

2. **Choose the procedure.** For each scenario, name the procedure — χ^2 **test of independence (or homogeneity)**, χ^2 **goodness-of-fit**, **two-proportion** z, or **not a** χ^2 **problem**.
   a. Is political party (3 levels) related to opinion on a policy (favor / oppose / undecided)?
   b. Is the proportion of "favor" responses different between Democrats and Republicans?
   c. Does the distribution of blood types (A/B/AB/O) in a sample of 500 match the population frequencies?
   d. Is the mean reaction time the same across three age groups?
   e. Is there an association between meal-plan choice (4 levels) and dorm (5 levels)?

3. **Check conditions before you trust the test.** Open the Chi-Square test tool with any two-categorical-variable dataset (e.g., bundled examples on relationships).
   a. What is the **expected-count condition** for the χ^2 independence test? Why are expected counts (not observed) the criterion?
   b. How are the **expected counts** in cell (i, j) computed? Walk through the formula in words.
   c. Suppose the smallest expected count is 4. Should you trust the χ^2 approximation? Name two responses that are more defensible than just running the test anyway.

4. **Run it and interpret it.** Open the Chi-Square test tool with a two-categorical dataset.
   a. Report the χ^2 statistic, df, and p-value.
   b. State your decision at $\alpha = 0.05$ and write a one-sentence conclusion **in context**.
   c. The tool reports the observed and expected counts side-by-side. Suppose H_0 is rejected — how do you decide *where* the association is strongest? What does a large positive $O - E$ in one cell mean substantively?
   d. A student concludes from a rejected H_0: "Major *causes* a student to choose a certain caffeine source." Why is this the wrong conclusion — and what *is* the right one?

5. **Simulation as a sanity check.** Open the Randomization χ^2 test on the same data.
   a. Run 5000 simulated χ^2 statistics. Compare its **simulation p-value** to the χ^2**-distribution p-value** from Exercise 4. How close are they?
   b. Repeat with a smaller dataset (or one cell deliberately small). Are the two p-values still close? Which would you trust?
   c. In a 2×2 table, the χ^2 statistic equals the square of the two-proportion z statistic, so the χ^2 test of independence gives the *same* p-value as a two-sided two-proportion z-test. Why is this not a coincidence?

d. Suppose you have a 3×3 table and H_0 is rejected. Why is "the variables are associated" all you can conclude from the test alone? What additional analyses would help pinpoint *which* groups differ?

Chapter 17

Analysis of Variance

In earlier chapters, we developed methods to compare the mean of a quantitative outcome across two groups. An important aspect of those analyses was to look at the difference in sample means as an estimate for the difference in population means. When comparing more than two groups, simple subtraction will not fully capture the variation across three or more groups. As with two groups, the research question focuses on whether group membership is independent of the quantitative response variable. In this chapter, we focus on a new statistic that incorporates differences in means across more than two groups. Although the ideas are similar to the t-test, they have earned their own name: **AN**alysis **O**f **VA**riance, or **ANOVA**.

Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons. For example, if there were three groups, we might compare the first mean with the second, then with the third, and finally compare the second and third means – a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually find a difference just by chance, even if there is no difference in the populations. Instead, we should apply a holistic test to check whether there is evidence that at least one pair of groups is different, which is where ANOVA saves the day.

ANOVA uses a single hypothesis test to check whether the means across many groups are equal:

- H_0: The mean outcome is the same across all groups: $\mu_1 = \mu_2 = \cdots = \mu_k$ where μ_j represents the mean of the outcome for observations in category j.
- H_A: At least one mean is different.

Generally, we must check three conditions before performing ANOVA:

- The observations are independent within and between groups.
- The responses within each group are nearly normal.
- The variability across the groups is about equal.

When these conditions are met, we may perform ANOVA to determine whether the data provide convincing evidence against the null hypothesis.

College departments commonly run multiple sections of the same introductory course. Consider a statistics department that runs three sections of an intro stats course. We might like to determine whether there are substantial differences in first exam scores across sections A, B, and C. Describe appropriate hypotheses.

- H_0: The average score is identical in all sections, $\mu_A = \mu_B = \mu_C$. The observed differences are due to chance.
- H_A: The average score varies by class. We would reject the null hypothesis if there were larger differences among the class averages than what we might expect from chance alone.

Strong evidence favoring the alternative hypothesis in ANOVA is described by unusually large differences among the group means. Assessing the variability of the group means relative to the variability among individual observations within each group is key to ANOVA's success.

Consider the figure below showing two sets of three groups.

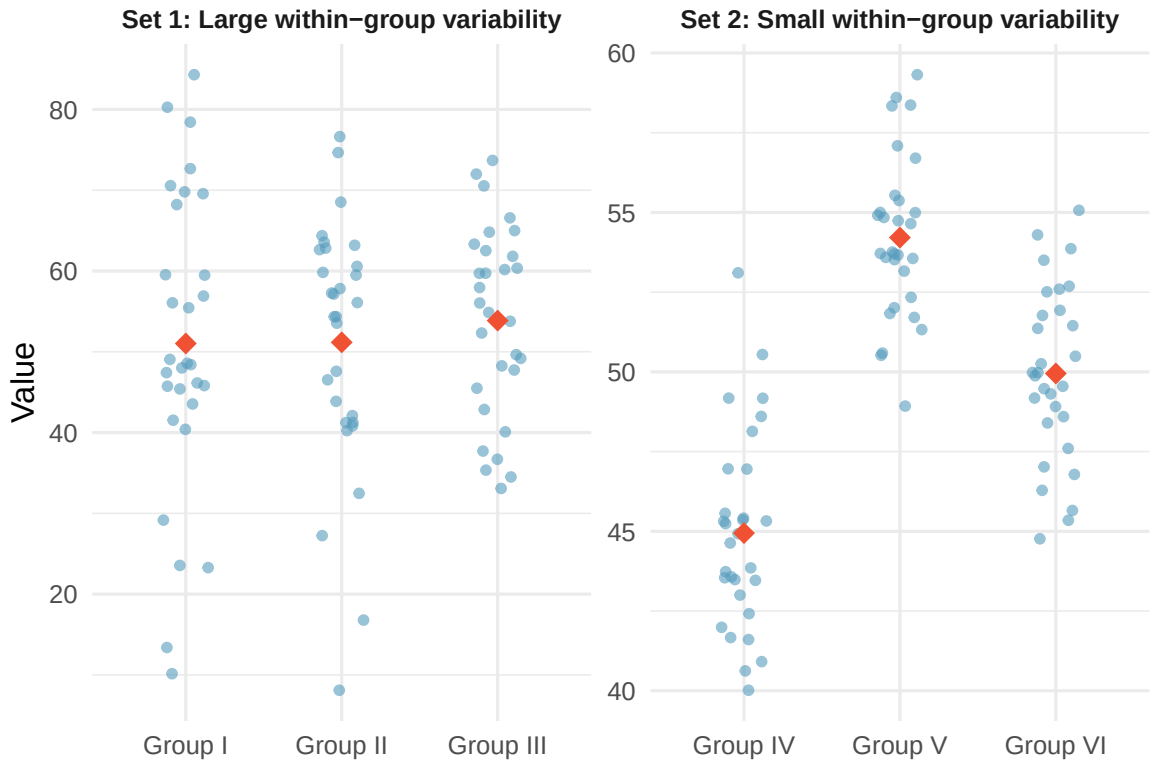


Figure 17.1: Side-by-side dot plots for two sets of three groups. In the first set (Groups I, II, III), the within-group variability is large relative to the between-group differences. In the second set (Groups IV, V, VI), the within-group variability is small and the between-group differences are clearly visible.

Compare Groups I, II, and III. Can you visually determine if the group differences are real? Now compare Groups IV, V, and VI.

Any real difference in Groups I, II, and III is difficult to discern because the data within each group are very volatile relative to differences in the group averages. On the other hand, Groups IV, V, and VI show noticeable differences because those differences are large *relative to the variability in individual observations within each group*.

17.1 Case study: MLB batting

We would like to determine whether there are real differences in batting performance among baseball players according to their position: outfielder (OF), infielder (IF), and catcher (C). We use data from 429 Major League Baseball players from the 2018 season who had at least 100 at-bats. The measure of batting performance is the on-base percentage (OBP), which roughly represents the fraction of the time a player successfully gets on base or hits a home run.

Sample statistics for each group:

Position	n	Mean	SD
OF	160	0.320	0.043
IF	205	0.318	0.038
C	64	0.302	0.038

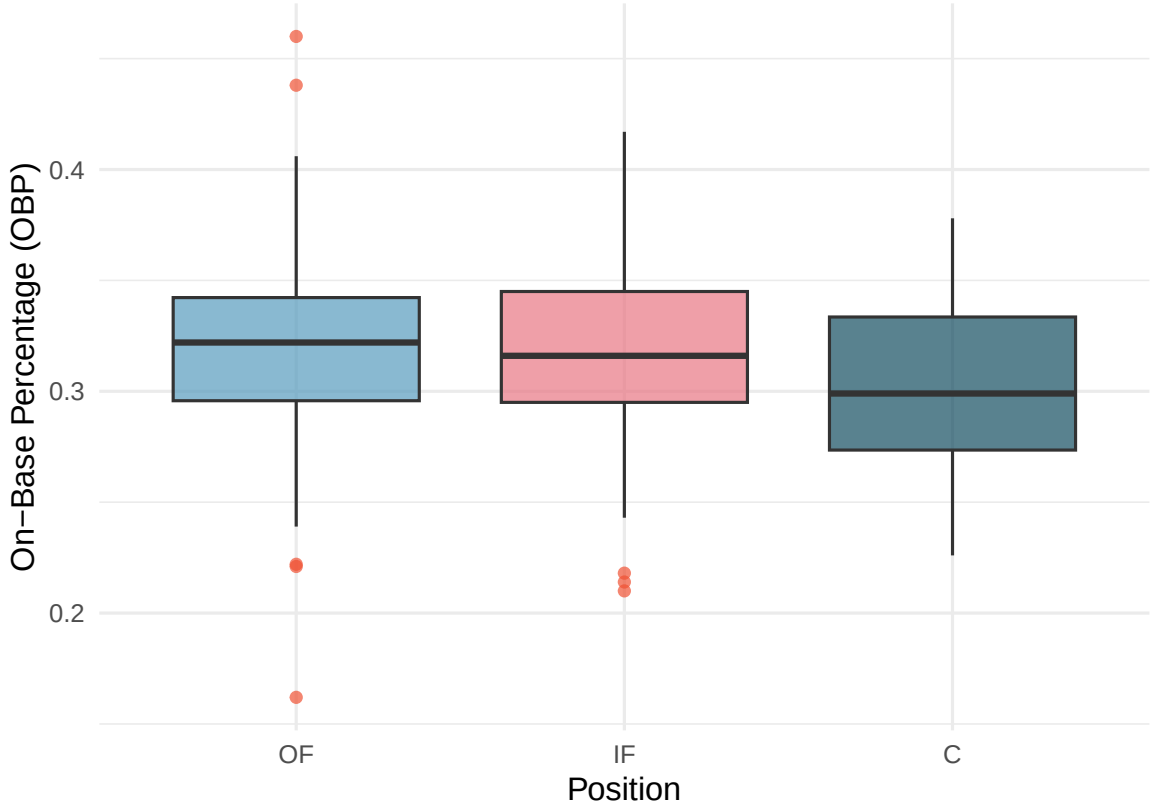


Figure 17.2: Side-by-side box plots of on-base percentage for 429 players across three positions. The variability appears approximately constant across groups. Catchers seem to have a slightly lower on-base percentage.

The null hypothesis under consideration is $\mu_{OF} = \mu_{IF} = \mu_C$. Write the null and alternative hypotheses in plain language.

Show answer

H_0: The average on-base percentage is equal across the three positions. H_A: The average on-base percentage varies across some (or all) groups.

What would be an appropriate point estimate of the on-base percentage by outfielders, μ_{OF}?

A good estimate of the on-base percentage by outfielders would be the sample average of OBP for just those players who are outfielders: $\bar{x}_{OF} = 0.320$.

17.1.1 Why not just compare the most different groups?

The largest difference between sample means is between catchers and outfielders. Why might it be inappropriate to simply test whether μ_C and μ_{OF} are different?

The primary issue is **data snooping** (also called **data fishing**). We are inspecting the data before picking the groups to compare. It is inappropriate to examine all data informally and only afterward decide which parts to formally test. Naturally, we would pick the groups with the largest differences, which inflates the Type I error rate.

For example, suppose we measure aptitude for students in 20 classrooms where all students are randomly assigned. With so many groups, we will probably observe a few that look different just by chance. If we select only those classes and perform a formal test, we will probably reach the wrong conclusion.

17.2 Randomization test for comparing many means

17.2.1 The F-statistic

The method of analysis of variance focuses on answering one question: is the variability in the sample means so large that it seems unlikely to be from chance alone? We quantify the between-group variability using the **mean square between groups** (**MSG**), which has degrees of freedom $df_G = k - 1$ where k is the number of groups. The MSG can be thought of as a scaled variance formula for the group means.

We also need a benchmark for how much variability to expect if the null hypothesis is true. For this, we compute the **mean square error** (**MSE**), which measures the variability *within* groups and has degrees of freedom $df_E = n - k$.

The **F-statistic** is the ratio of these two quantities:

$$F = \frac{MSG}{MSE}$$

The MSG represents between-group variability, and MSE measures within-group variability. When the null hypothesis is true, any differences among sample means are due to chance, and the MSG and MSE should be about equal, making F close to 1. Large values of F suggest that the between-group differences are larger than what we would expect from random variation.

The test statistic for three or more means is an F.

The F statistic is a ratio of how the groups differ (MSG) as compared to how the observations within a group vary (MSE).

$$F = \frac{MSG}{MSE}$$

When the null hypothesis is true and the conditions are met, F has an F-distribution with $df_1 = k-1$ and $df_2 = n - k$.

Conditions:

- Independent observations, both within and across groups
- Large samples and no extreme outliers

17.2.2 Variability of the statistic

Suppose we have three exams (A, B, and C) and want to investigate whether they differ in difficulty:

- H_0: $\mu_A = \mu_B = \mu_C$. The exams are equally difficult.
- H_A: At least one exam is more (or less) difficult than the others.

Exam	n	Mean	SD	Min	Max
A	58	75.1	13.9	44	100
B	55	72.0	13.8	38	100
C	51	78.9	13.1	45	100

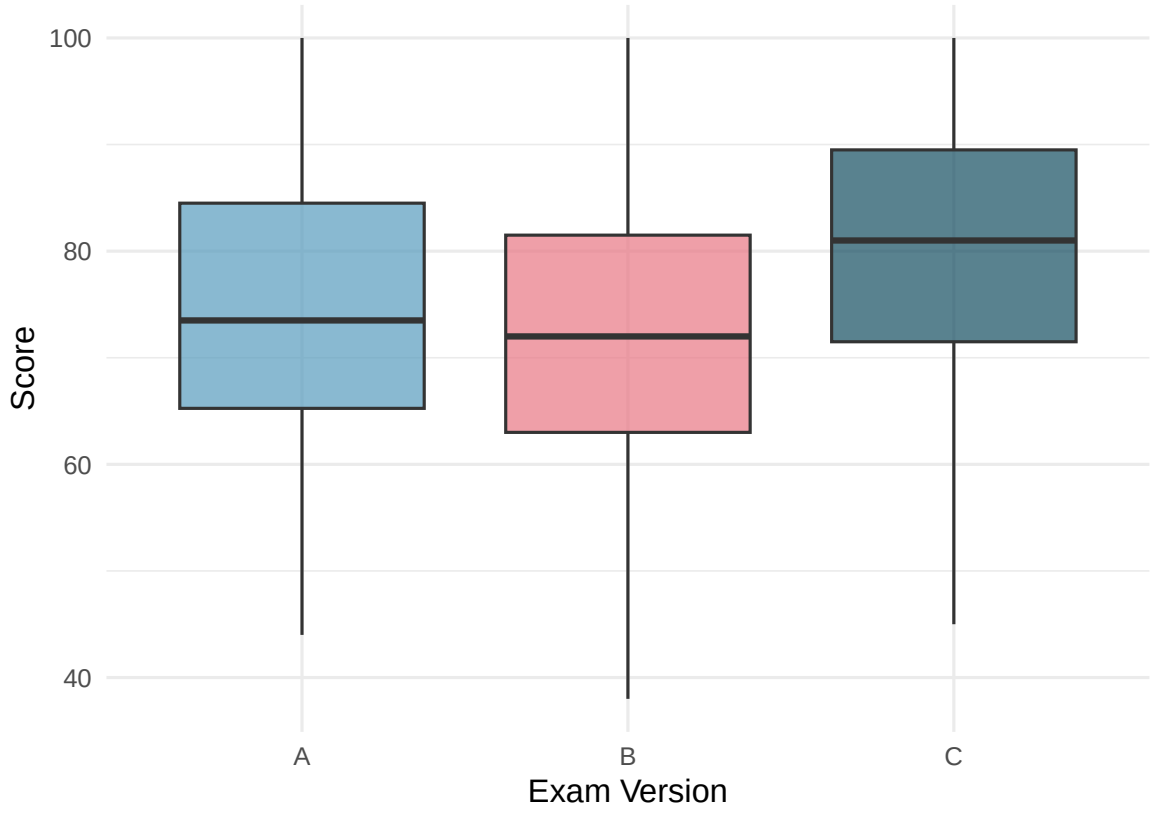


Figure 17.3: Box plots of exam scores for three exam versions. Exam C appears to have a higher median, but the distributions overlap.

To perform a randomization test, we randomly reassign the exam labels to the observed scores. If the null hypothesis is true, the exam version should not matter, so any score could belong to any exam version.

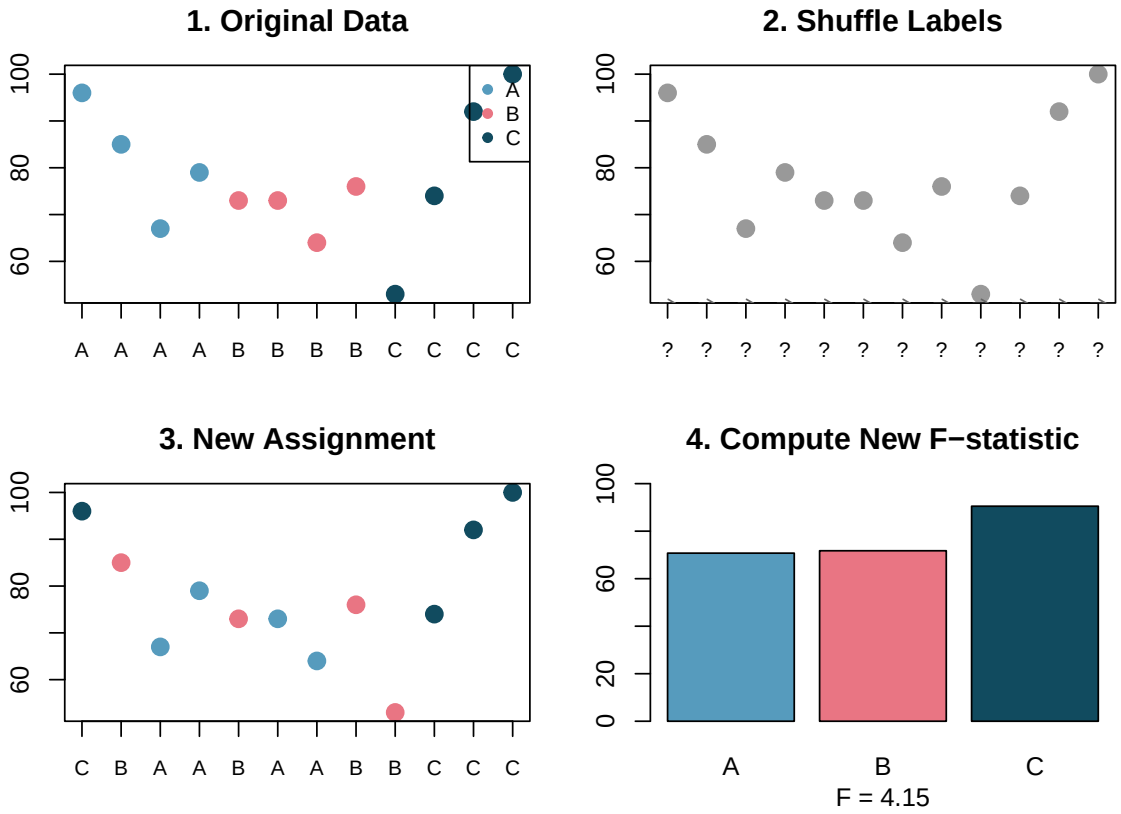


Figure 17.4: The randomization process, shown for 12 of the observed scores (4 from each exam version): exam version labels are randomly shuffled across the observed scores to generate a new assignment, from which a new F-statistic is computed. ↗ Try this live — shuffle the group labels yourself and see the resulting F-statistic.

By repeating this process 1,000 times, we build a null distribution of F-statistics.

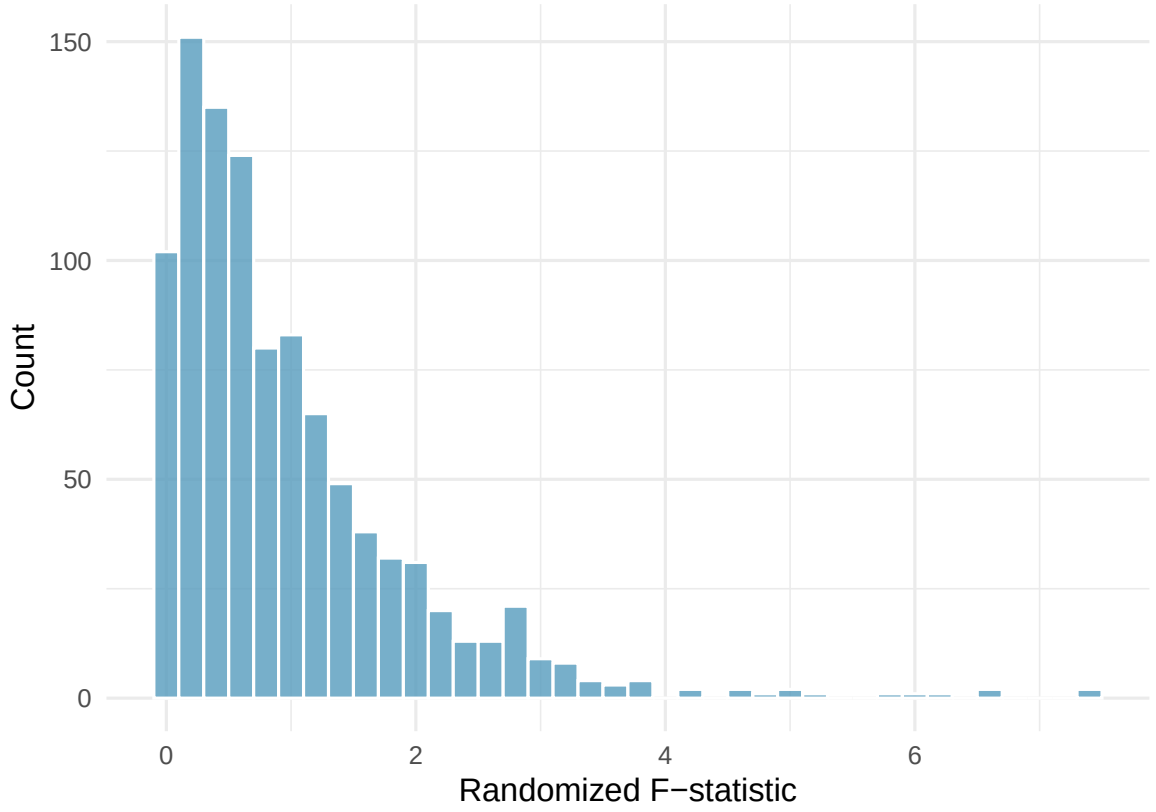


Figure 17.5: Histogram of F-statistics from 1,000 randomizations of the exam data. The distribution is right-skewed with most values below 2 and a long right tail. ↗ Try this live — build the null F distribution live — vary reps and watch it settle.

17.2.3 Observed statistic vs. null statistics

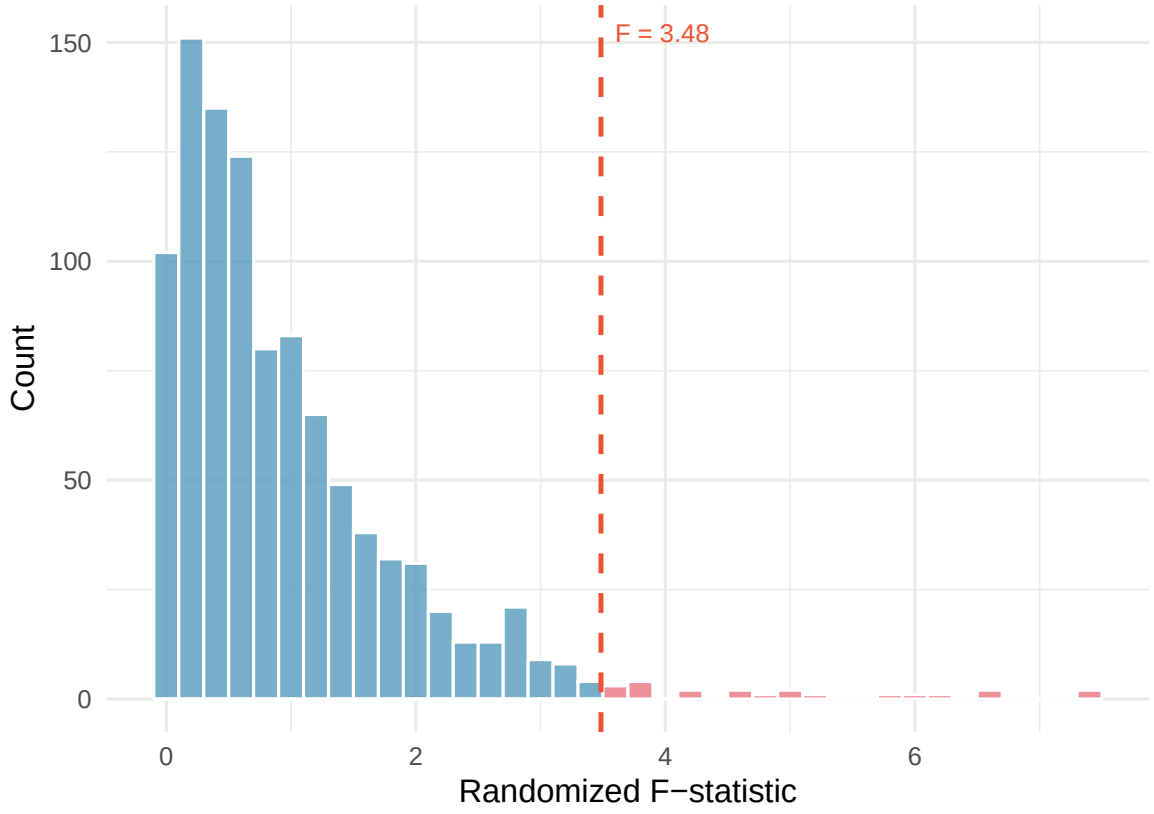


Figure 17.6: Histogram of F-statistics from 1,000 randomizations with the observed F-statistic of 3.48 marked by a red line. The area to the right represents the p-value. ↗ Try this live — reproduce the null distribution and place your observed F on it.

Using statistical software, we find that 2.2% of the randomized F statistics were at or above the observed test statistic of $F = 3.48$. That is, the p-value is 0.022. Assuming a discernibility level of $\alpha = 0.05$, the p-value is smaller than the discernibility level, which leads us to reject the null hypothesis. We conclude that the difficulty level is different for at least one of the exams.

17.3 Mathematical model for ANOVA

The randomization test on the F-statistic has a corresponding mathematical theory: the **F-distribution**.

17.3.1 Variability of the statistic

The larger the observed variability in the sample means (MSG) relative to the within-group variability (MSE), the larger the F-statistic and the stronger the evidence against the null hypothesis. We use the upper tail of the F-distribution to compute the p-value.

The F-statistic and the F-test.

Analysis of variance (ANOVA) uses a test statistic, the F-statistic, which represents a standardized ratio of variability in the sample means relative to the variability within the groups. If H_0 is true and

the model conditions are satisfied, the F-statistic follows an F-distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$. The upper tail of the F-distribution is used to compute the p-value.

For the MLB data, $MSG = 0.00803$ and $MSE = 0.00158$. Identify the degrees of freedom associated with MSG and MSE and verify the F-statistic is approximately 5.077.

Show answer

There are $k = 3$ groups, so $df_G = k - 1 = 2$. There are $n = 429$ total observations, so $df_E = n - k = 426$. The F-statistic is $F = \frac{MSG}{MSE} = \frac{0.00803}{0.00158} = 5.08 \approx 5.077$.

17.3.2 The p-value from the F-distribution

A p-value can be computed from the F-statistic using an F-distribution with $df_1 = df_G$ and $df_2 = df_E$.

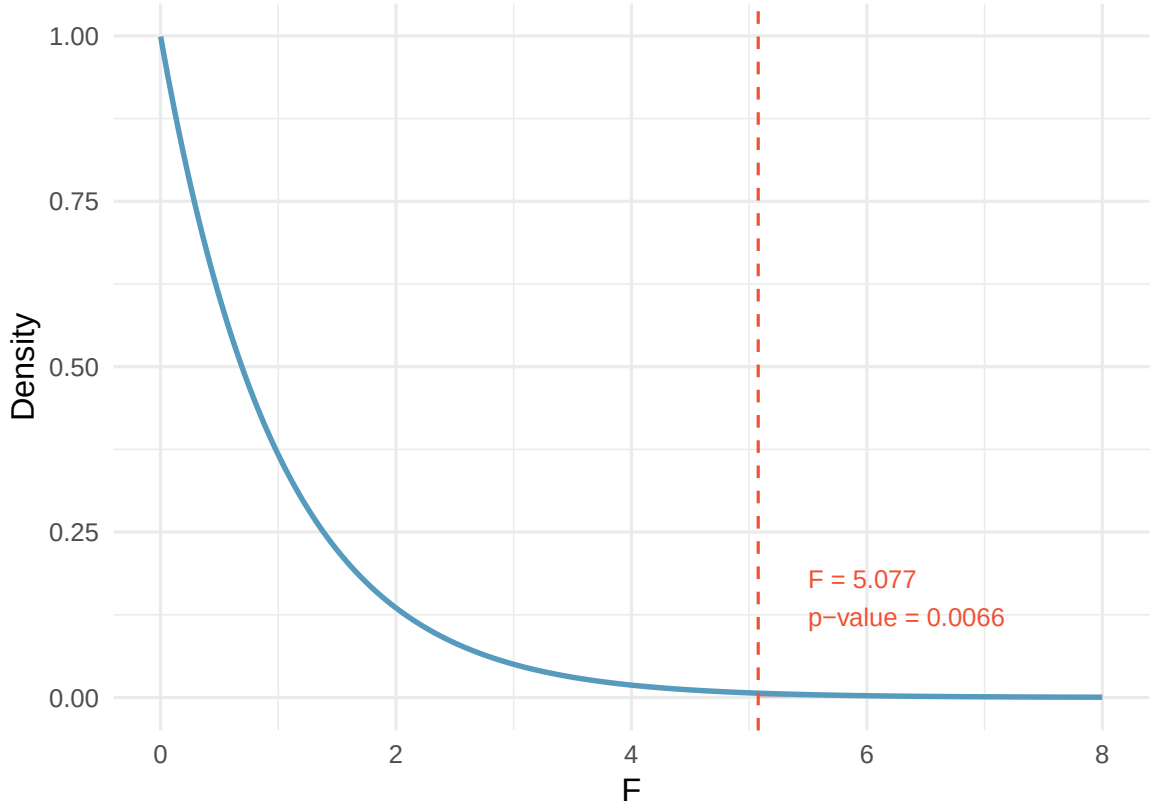


Figure 17.7: An F-distribution with df□ = 2 and df□ = 426. The area to the right of F = 5.077 is shaded, representing a p-value of approximately 0.0066. ↗ Try this live — drag the cutline to any F value and read the right-tail p-value.

The p-value corresponding to $F = 5.077$ with $df_1 = 2$ and $df_2 = 426$ is approximately 0.0066. Does this provide strong evidence against the null hypothesis?

The p-value is smaller than 0.05, indicating the evidence is strong enough to reject the null hypothesis at a discernibility level of 0.05. The data provide strong evidence that the average on-base percentage varies by player's primary field position.

Note that the small p-value tells us there is a notable difference somewhere, but ANOVA does **not** tell us *which* group is different. Follow-up analyses (discussed in the next chapter on multiple comparisons) are needed.

17.3.3 Reading an ANOVA table

ANOVA results are commonly presented in a table. Here is the ANOVA table for the MLB data:

Source	df	Sum Sq	Mean Sq	F value	p-value
position	2	0.0161	0.00803	5.077	0.0066
Residuals	426	0.6740	0.00158		

The key information is in the last two columns: the F-statistic (5.077) and the p-value (0.0066).

Components of an ANOVA table.

- **Sum of Squares Between Groups (SSG):** Measures the total variability of the group means around the overall mean. $SSG = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2$.
- **Sum of Squared Errors (SSE):** Measures the total variability of observations around their group means. $SSE = \sum_{j=1}^{k}(n_j - 1)s_j^2$.
- **Sum of Squares Total (SST):** The total variability in all observations. $SST = SSG + SSE$.
- **Mean Square Between Groups (MSG):** $MSG = SSG/(k-1)$.
- **Mean Square Error (MSE):** $MSE = SSE/(n-k)$.
- **F-statistic:** $F = MSG/MSE$.

See it in action. Open the ANOVA omnibus walkthrough for a guided pass through framing → conditions → running the F → the "at least one differs" caveat. Or open the One-Way ANOVA tool or Randomization ANOVA directly. **For post-hoc pairwise comparisons**, append `?posthoc=tukey` (or `bonferroni`) — see Ch 18.

17.4 Conditions for ANOVA

There are three conditions to check before performing ANOVA:

- **Independence.** If the data are a simple random sample, this condition is generally satisfied. For experiments, carefully consider whether the data may be independent (e.g., no pairing).
- **Approximately normal.** As with one- and two-sample testing for means, the normality assumption is especially important when sample sizes are small. With large sample sizes within each group, we look mainly for extreme outliers.
- **Constant variance.** The variability in the groups should be about equal. This can be checked by examining side-by-side box plots or comparing the standard deviations across groups. A common rule of thumb is that the ratio of the largest to smallest group standard deviation should be less than 2.

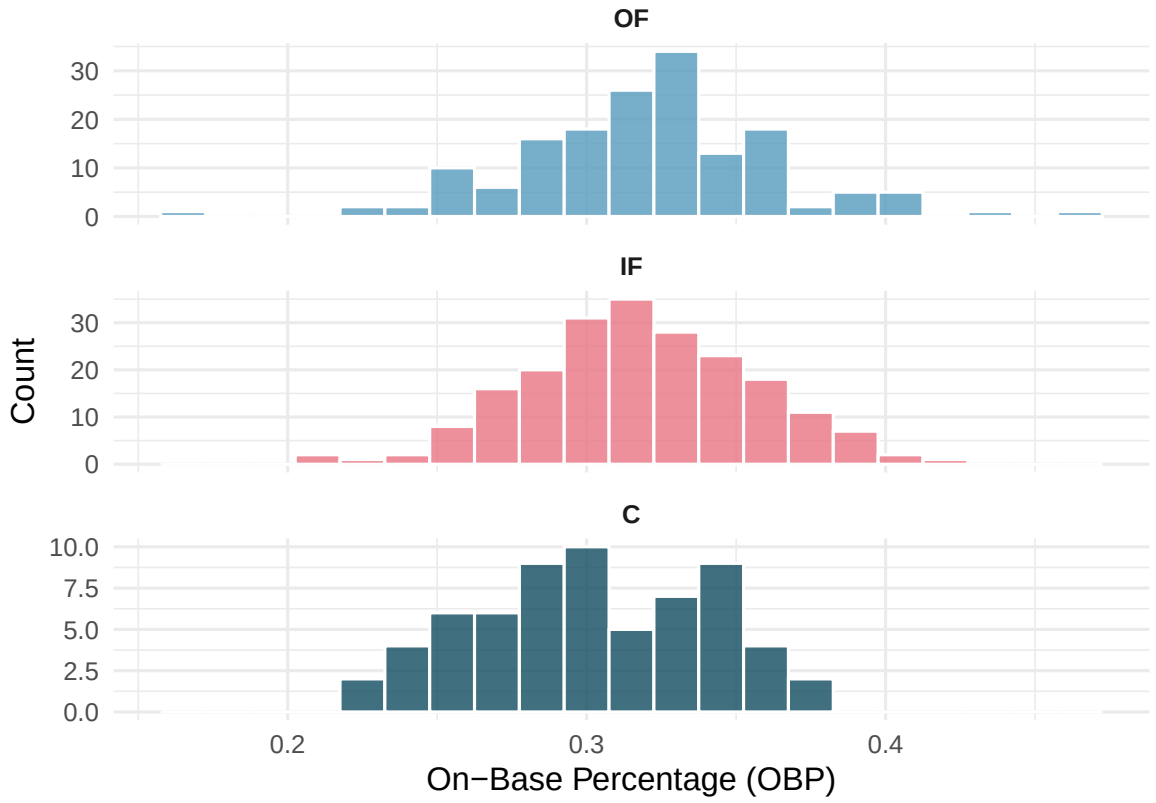


Figure 17.8: Histograms of OBP for each field position, showing that the distributions are reasonably bell-shaped and symmetric.

Diagnostics for an ANOVA analysis.

Independence is always important. The normality condition matters most when group sizes are small (with large groups, look only for extreme outliers). The constant variance condition is especially important when group sizes differ.

17.5 Chapter review

17.5.1 Summary

In this chapter, we introduced both the randomization test and the mathematical model for comparing means across two or more groups using ANOVA. The F-statistic measures the ratio of between-group variability (MSG) to within-group variability (MSE). When the null hypothesis is true and the conditions are satisfied, the F-statistic follows an F-distribution. A discernible result tells us that at least one group mean differs from the others, but does not identify *which* groups differ. For that, we need multiple comparison procedures — Bonferroni-adjusted pairwise CIs, Tukey's HSD, and related methods — covered in the multiple comparisons chapter.

17.5.2 Key terms

- ANOVA (analysis of variance)
- F-statistic
- F-distribution

- F-test
- Mean square between groups (MSG)
- Mean square error (MSE)
- Sum of squares between groups (SSG)
- Sum of squared errors (SSE)
- Sum of squares total (SST)
- Degrees of freedom (ANOVA)
- Data snooping / data fishing

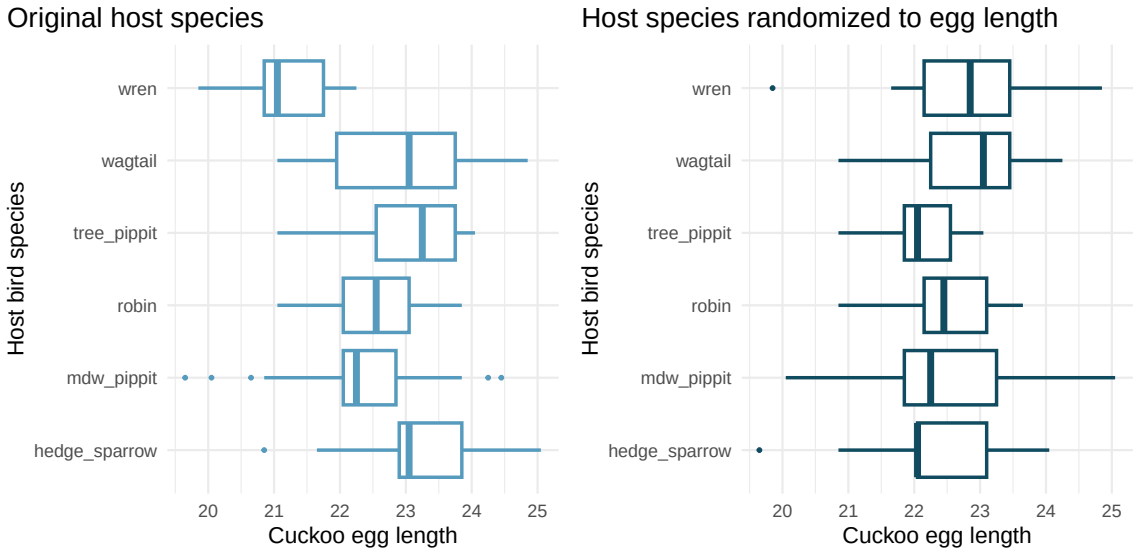
17.6 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Fill in the blank.** When doing an ANOVA, you observe large differences in means between groups. Within the ANOVA framework, this would most likely be interpreted as evidence strongly favoring the _____ hypothesis.

2. **Which test?** We would like to test if students who are in the social sciences, natural sciences, arts and humanities, and other fields spend the same amount of time, on average, studying for a course. What type of test should we use? Explain your reasoning.

3. **Cuckoo bird egg lengths, randomize once.** Cuckoo birds lay their eggs in other birds' nests, making them known as brood parasites. One question relates to whether the size of the cuckoo egg differs depending on the species of the host bird. (Latter 1902) Consider the following plots, one represents the original data, the second represents data where the host species has been randomly assigned to the egg length.

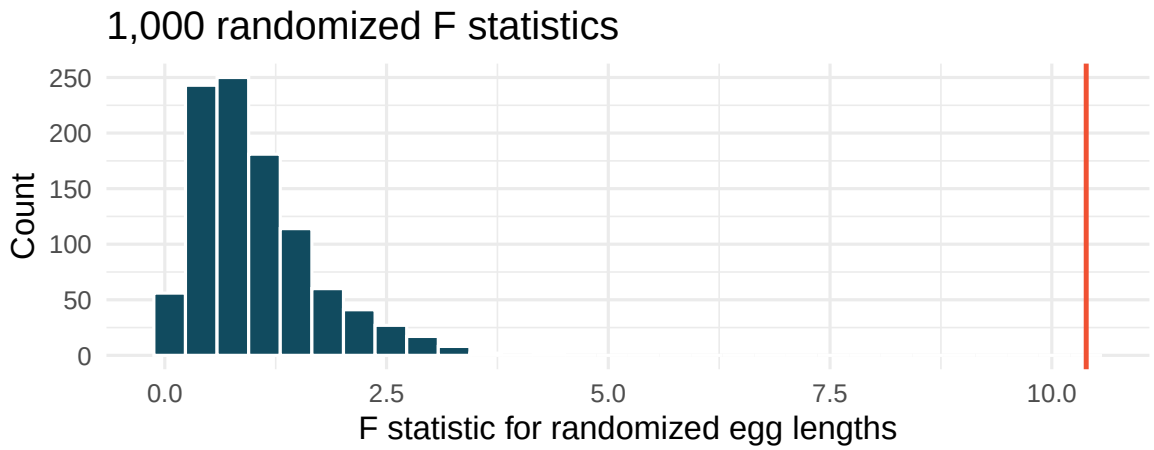


a. Consider the average length of the eggs for each species. Is the average length for the original data: more variable, less variable, or about the same as the randomized species? Describe what you see in the plots.

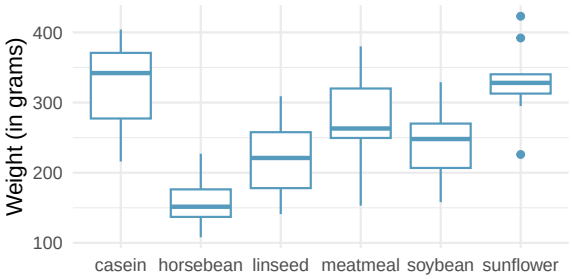
b. Consider the standard deviation of the lengths of the eggs within each species. Is the within species standard deviation of the length for the original data: bigger, smaller, or about the same as the randomized species?

c. Recall that the F statistic's numerator measures how much the groups vary (MSG) with the denominator measuring how much the within species values vary (MSE), which of the plots above would have a larger F statistic, the original data or the randomized data? Explain.

4. **Cuckoo bird egg lengths, randomization test.** Cuckoo birds lay their eggs in other birds' nests, making them known as brood parasites. One question relates to whether the size of the cuckoo egg differs depending on the species of the host bird. (Latter 1902) Using the randomization distribution of the F statistic (host species randomized to egg length), conduct a hypothesis test to evaluate if there is a difference, in the population, between the average egg lengths for different host bird species. Make sure to state your hypotheses clearly and interpret your results in context of the data.



5. **Chicken diet and weight, many groups.** An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Sample statistics and a visualization of the observed data are shown below. (McNeil 1977)



Feed type	Mean	SD	n
casein	323.58	64.43	12
horsebean	160.20	38.63	10
linseed	218.75	52.24	12
meatmeal	276.91	64.90	11
soybean	246.43	54.13	14
sunflower	328.92	48.84	12

Using the ANOVA output below, conduct a hypothesis test to determine if these data provide convincing evidence that the average weight of chicks varies across some (or all) groups. Make sure to check relevant conditions.

term	df	sumsq	meansq	statistic	p.value
`feed`	5	231,129	46,226	15.4	<0.0001
`Residuals`	65	195,556	3,009	NA	NA

6. **Teaching descriptive statistics.** A study compared five different methods for teaching descriptive statistics. The five methods were traditional lecture and discussion, programmed textbook instruction, programmed text with lectures, computer instruction, and computer instruction with lectures. 45 students were randomly assigned, 9 to each method. After completing the course, students took a 1-hour exam.

a. What are the hypotheses for evaluating if the average test scores are different for the different teaching methods?

b. What are the degrees of freedom associated with the F-test for evaluating these hypotheses?

c. Suppose the p-value for this test is 0.0168. What is the conclusion?

7. **Coffee, depression, and physical activity.** Caffeine is the world's most widely used stimulant, with approximately 80% consumed in the form of coffee. Participants in a study investigating the relationship between coffee consumption and exercise were asked to report the number of hours they spent per week on moderate (e.g., brisk walking) and vigorous (e.g., strenuous sports and jogging) exercise. Based on these data the researchers estimated the total hours of metabolic equivalent tasks (MET) per week, a value always greater than 0. The table below gives summary statistics of MET for women in this study based on the amount of coffee consumed. (Lucas et al. 2011)

	Caffeinated coffee consumption				
	1 cup / week or fewer	2-6 cups / week	1 cups / day	2-3 cups / day	4 cups / day or more
Mean	18.7	19.6	19.3	18.9	17.5
SD	21.1	25.5	22.5	22.0	22.0
n	12,215.0	6,617.0	17,234.0	12,290.0	2,383.0

a. Write the hypotheses for evaluating if the average physical activity level varies among the different levels of coffee consumption.

b. Check conditions and describe any assumptions you must make to proceed with the test.

c. Below is the output associated with this test. What is the conclusion of the test?

	df	sumsq	meansq	statistic	p.value
`cofee`	4	10,508	2,627	5.2	0
`Residuals`	50,734	25,564,819	504	NA	NA
`Total`	50,738	25,575,327	NA	NA	NA

8. **Student performance across discussion sections.** A professor who teaches a large introductory statistics class (197 students) with eight discussion sections would like to test if student performance differs by discussion section, where each discussion section has a different teaching assistant. The summary table below shows the average final exam score for each discussion section as well as the standard deviation of scores and the number of students in each section.

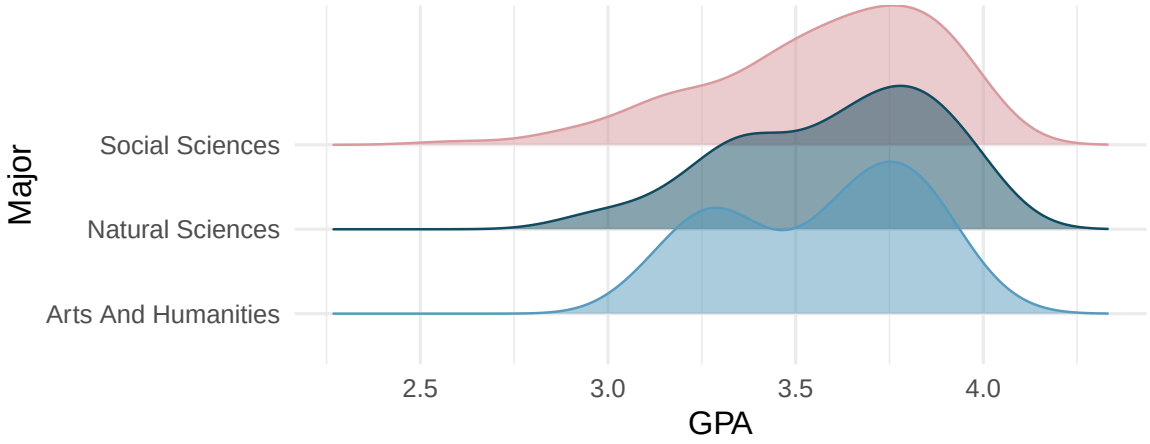
	Sec 1	Sec 2	Sec 3	Sec 4	Sec 5	Sec 6	Sec 7	Sec 8
Mean	92.94	91.11	91.80	92.45	89.30	88.30	90.12	93.35
SD	4.21	5.58	3.43	5.92	9.32	7.27	6.93	4.57
n	33.00	19.00	10.00	29.00	33.00	10.00	32.00	31.00

The ANOVA output below can be used to test for differences between the average scores from the different discussion sections.

	df	sumsq	meansq	statistic	p.value
`section`	7	525	75.0	1.87	0.077
`Residuals`	189	7,584	40.1	NA	NA
`Total`	196	8,109	NA	NA	NA

Conduct a hypothesis test to determine if these data provide convincing evidence that the average score varies across some (or all) groups. Check conditions and describe any assumptions you must make to proceed with the test.

9. **GPA and major.** Undergraduate students in an introductory statistics course at Duke University conducted a survey about GPA and major. The density plots show the distributions of GPA among three groups of majors. ANOVA output is also provided.



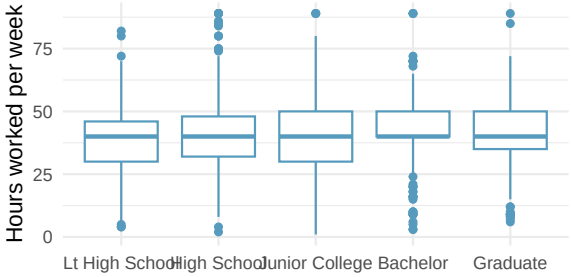
term	df	sumsq	meansq	statistic	p.value
`major`	2	0.03	0.02	0.21	0.81
`Residuals`	195	15.77	0.08	NA	NA

a. Write the hypotheses for testing for a difference between average GPA across majors.

b. What is the conclusion of the hypothesis test?

c. How many students answered the questions on the survey, i.e., what is the sample size?

10. **Work hours and education.** The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents. (NORC 2010) Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.



Educational attainment	Mean	SD	n
Lt High School	38.7	15.8	121
High School	39.6	15.0	546
Junior College	41.4	18.1	97
Bachelor	42.5	13.6	253
Graduate	40.8	15.5	155

a. Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.

b. Check conditions and describe any assumptions you must make to proceed with the test.

c. Below is the output associated with this test. What is the conclusion of the test?

term	df	sumsq	meansq	statistic	p.value
degree	4	2,006	502	2.19	0.07
Residuals	1,167	267,382	229	NA	NA

11. **Matching observed data with randomized F statistics.** Consider the following two datasets. The response variable is the `score` and the explanatory variable is whether the individual is in one of four groups.

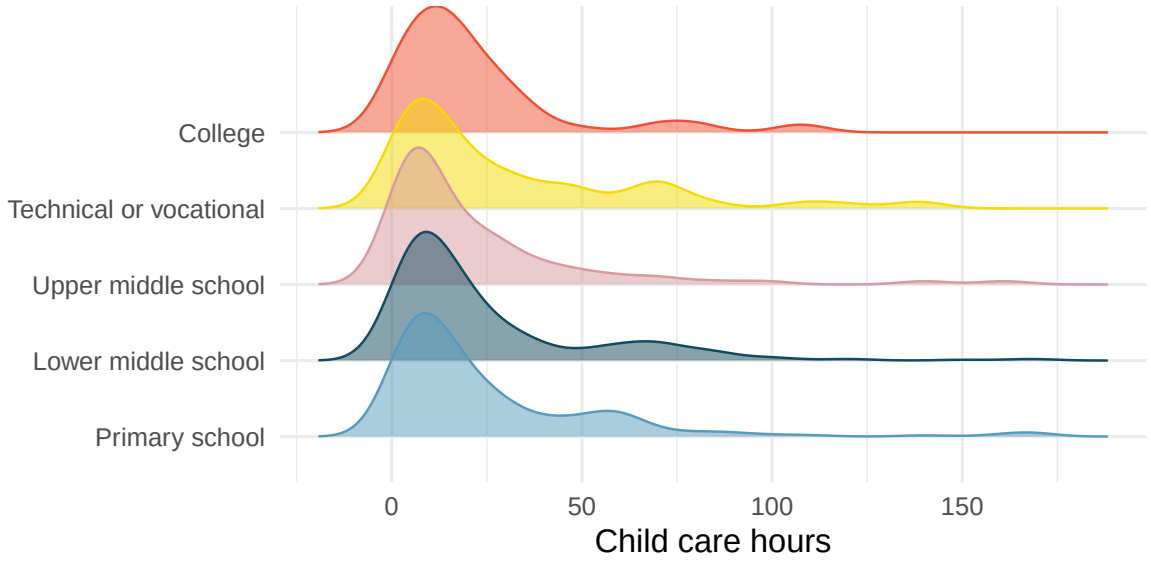


The randomizations (randomly assigning group to the score, calculating a randomization F statistic) were done 1000 times for each of Dataset A and B. The red line on each plot indicates the observed F statistic for the original (unrandomized) data.

a. Does the randomization distribution on the left correspond to Dataset A or B? Explain.

b. Does the randomization distribution on the right correspond to Dataset A or B? Explain.

12. **Child care hours.** The China Health and Nutrition Survey aims to examine the effects of the health, nutrition, and family planning policies and programs implemented by national and local governments. (UNC Carolina Population Center 2006) It, for example, collects information on number of hours Chinese parents spend taking care of their children under age 6. The side-by-side box plots below show the distribution of this variable by educational attainment of the parent. Also provided below is the ANOVA output for comparing average hours across educational attainment categories.



term	df	sumsq	meansq	statistic	p.value
`edu`	4	4,142	1,036	1.26	0.28
`Residuals`	794	653,048	822	NA	NA

a. Write the hypotheses for testing for a difference between the average number of hours spent on child care across educational attainment levels.

b. What is the conclusion of the hypothesis test?

13. **After a significant ANOVA, then what?** A one-way ANOVA compares mean commute times across 4 city types (rural, suburban, small-city, big-city) using data from $n = 200$ commuters. The ANOVA rejects H_0: all population means are equal ($F = 8.4$, p-value $= 0.00003$).

a. What does the significant F-test tell you? What does it *not* tell you?

b. A colleague suggests building all $\binom{4}{2} = 6$ pairwise 95% CIs for the differences in means (rural − suburban, rural − small-city, rural − big-city, suburban − small-city, suburban − big-city, small-city − big-city). If each CI is constructed at the 95% confidence level, will the *joint* coverage — the probability that all six CIs simultaneously capture their true differences — be 95%? Above 95%? Below 95%? Explain in one sentence.

c. Name one method (from Ch 18 in this textbook, or from other courses) that adjusts pairwise CIs so that the joint coverage stays at (or above) 95%.

d. Why do we care about joint coverage rather than just individual coverage?

Datasets: `Cuckoo` (Stat2Data)

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Frame the question.** An agronomist randomly assigns plots to one of **four** fertilizer brands and measures yield (bushels/acre) at harvest. She has 30 plots per brand.
   a. Identify the explanatory variable, its number of levels, and the response.
   b. Write H_0 and H_A for the omnibus ANOVA F-test in symbols *and* in words.
   c. Why does ANOVA use an F-statistic that compares two *variances*, when the question is really about *means*?

2. **Choose the procedure.** For each scenario, name the procedure — **one-way ANOVA F-test**, **two-sample t-test**, **paired t-test**, or **not an ANOVA problem**.
   a. Compare mean test scores across three teaching methods (random assignment).
   b. Compare mean test scores between two teaching methods (random assignment).
   c. Compare the *proportion* of students passing across four schools.
   d. Compare yields across four fertilizer brands, where the *same* plots received different brands in different years (within-plot pairing).
   e. Compare the mean grade of every student in three sections of the same course.

3. **Check the conditions.** Open the One-Way ANOVA tool with any multi-group numeric dataset.
   a. Name the **three** ANOVA conditions and state each in plain language.
   b. Inspect the side-by-side boxplots in StatLens. What pattern would convince you the **equal-variance** condition has failed?
   c. Suppose one group has $n_1 = 8$ with a strong right skew, while two other groups have $n_2 = n_3 = 50$ and look symmetric. Is ANOVA's F-test trustworthy here? What's a defensible alternative?

4. **Run it and interpret it.** Open the One-Way ANOVA tool with any multi-group numeric dataset.
   a. Report the F-statistic, the two df values (between and within), and the p-value.
   b. State your decision at $\alpha = 0.05$ and write a one-sentence conclusion **in context**.
   c. Suppose F is large and the p-value is 0.002. The conclusion "the four groups all have different means" is **wrong**. Why? What is the correct conclusion?
   d. Which statement is the correct interpretation of $F = 4.5$?
      i. Group means differ by an average of 4.5 units.
      ii. The between-group variance is 4.5 times the within-group variance, in this sample.
      iii. About 4.5% of the variability is explained by the group factor.

5. **Simulation as a sanity check.** Open the Randomization ANOVA on the same data.
   a. Run 5000 simulated F statistics under H_0 (group labels shuffled). What does the null distribution **look like**? Is it F-shaped?
   b. Compare the **simulation** p-value to the **F-distribution** p-value from Exercise 4. How close are they?
   c. Repeat on a dataset where one group has only $n = 6$ observations with a clear outlier. Are the two p-values still close? Which is the safer one to report?
   d. A student says: "If I'm going to run the randomization version anyway, why learn the F-distribution?" Give two good reasons. (Hint: think about *understanding* the F statistic, and *speed* for routine reporting.)

Chapter 18

Multiple Comparisons

Extended Content. This chapter extends beyond a standard one-semester introductory course. It is included for completeness and for instructors who wish to cover additional topics. Content is in draft form.

When an ANOVA test rejects the null hypothesis, we know that at least one group mean differs from the others – but we do not know *which* groups differ. This chapter addresses the natural follow-up question: which specific pairs of groups have different means? Attempting many pairwise comparisons inflates the overall Type I error rate, so we need specialized methods that control for the fact that we are conducting multiple tests simultaneously. We introduce the multiple testing problem, the Bonferroni correction, and Tukey's Honest Significant Difference (HSD) method.

See it in action. For the multiple-testing problem itself, work through the story-driven Do Green Jelly Beans Cause Acne? activity (family-wise error inflation, built around xkcd #882). For the actual post-hoc procedure, open Multiple Comparisons on any 3+ group numeric dataset — pairwise-CI table + forest plot, toggle Tukey HSD ↔ Bonferroni. Try insectsprays for the classic clean pattern.

18.1 The multiple testing problem

When we conduct a single hypothesis test at discernibility level $\alpha = 0.05$, there is a 5% chance of incorrectly rejecting a true null hypothesis (a Type I error). But what happens when we conduct many tests?

Suppose we compare $k = 4$ groups. How many pairwise comparisons are possible? If we conduct each test at $\alpha = 0.05$ and all null hypotheses are actually true, what is the probability that we make at least one Type I error?

With 4 groups, there are $\binom{4}{2} = 6$ pairwise comparisons. If we conduct each at $\alpha = 0.05$, the probability that any *single* test does not produce a Type I error is $1-0.05 = 0.95$. If the tests were independent, the probability that *none* of the 6 tests produces a Type I error would be:

$$P(\text{no errors}) = 0.95^6 = 0.735$$

So the probability of making **at least one** Type I error is:

$$P(\text{at least one error}) = 1 - 0.735 = 0.265$$

That is about a 27% chance of at least one false positive if the tests were independent. Pairwise comparisons are not actually independent – each group appears in several comparisons – so the true chance is somewhat lower (about 20% for all six comparisons among four groups), but it is still much higher than the 5% we intended.

18.1.1 Family-wise error rate

The probability of making at least one Type I error across all tests in a family of comparisons is called the **family-wise error rate** (FWER). When conducting m independent tests, each at discernibility level α, the FWER is:

$$\text{FWER} = 1 - (1 - \alpha)^m$$

The table below applies this independent-tests formula to all pairwise comparisons among k groups when $\alpha = 0.05$:

Number of groups (k)	Number of comparisons (m)	FWER from the independent-tests formula
3	3	0.143
4	6	0.265
5	10	0.401
6	15	0.537
10	45	0.901

These values are an approximation. Pairwise comparisons share groups, so the tests are correlated, and the actual FWER for unadjusted pairwise t-tests is lower than the formula gives – roughly 0.12 for 3 groups and about 0.6 for 10 groups. Even so, with 10 groups we are more likely than not to make at least one Type I error if we use unadjusted pairwise tests. This is the core of the multiple comparisons problem: we need procedures that control the FWER at an acceptable level (typically 0.05).

Family-wise error rate (FWER).

The **family-wise error rate** is the probability of making one or more Type I errors among all the hypothesis tests in a family of comparisons. Methods like the Bonferroni correction and Tukey's HSD are designed to keep the FWER at or below a specified level α.

18.2 The Bonferroni correction

The simplest method for controlling the family-wise error rate is the **Bonferroni correction**. The idea is straightforward: if we are conducting m tests and want the overall FWER to be at most α, we use a stricter discernibility level for each individual test.

Bonferroni correction.

When conducting m hypothesis tests and we want the family-wise error rate to be at most α, the Bonferroni correction uses an adjusted discernibility level for each individual test:

$$\alpha^* = \frac{\alpha}{m}$$

Equivalently, we can multiply each p-value by m (the adjusted p-value) and compare the result to the original α.

After an ANOVA with $k = 4$ groups rejects H_0, we want to conduct all pairwise comparisons at an overall $\alpha = 0.05$. How many comparisons are there, and what adjusted discernibility level should we use for each?

There are $\binom{4}{2} = 6$ pairwise comparisons. The Bonferroni-adjusted discernibility level is:

$$\alpha^* = \frac{0.05}{6} = 0.0083$$

So each individual two-sample t-test must have a p-value below 0.0083 to be considered discernible.

The Bonferroni correction is simple and broadly applicable – it can be used with any type of test, not just pairwise comparisons of means. However, it is **conservative**, meaning it may fail to detect real differences when they exist (higher Type II error). When the number of comparisons is very large, the Bonferroni correction can be overly strict.

18.2.1 Example: Exam scores

The ANOVA comparing the three exam versions (A, B, C) in the previous chapter gave $F = 3.48$, with a p-value of 0.033 from the F-distribution. We rejected H_0 and now want to know which exams differ.

With $k = 3$ groups, there are $\binom{3}{2} = 3$ pairwise comparisons. The Bonferroni-adjusted discernibility level is $\alpha^* = 0.05/3 = 0.0167$.

Two-sample t-tests for the three pairs (using the pooled standard deviation from the ANOVA) give these p-values:

Comparison	p-value	Discernible at $\alpha^* = 0.0167$?
A vs. B	0.222	No
A vs. C	0.144	No
B vs. C	0.009	Yes

Only the comparison between exams B and C is discernible after the Bonferroni correction. We conclude that there is evidence of a difference in difficulty between exams B and C, but we cannot say the other pairs differ.

18.3 Tukey's Honest Significant Difference

A more powerful alternative to the Bonferroni correction for pairwise comparisons of means is **Tukey's Honest Significant Difference (HSD)** method. While the Bonferroni correction is general-purpose, Tukey's method is specifically designed for comparing all pairs of group means after ANOVA, and it is less conservative.

18.3.1 How Tukey's HSD works

Tukey's HSD computes a single critical difference – the minimum difference between two group means that would be considered statistically discernible. The procedure uses the **Studentized range distribution** to determine this critical value.

For each pair of groups j and j', the test statistic is:

$$q = \frac{\bar{x}_j - \bar{x}_{j'}}{\sqrt{MSE/n^*}}$$

where MSE is the mean square error from the ANOVA, and n^* is a measure of the group sample size (when group sizes are equal, $n^* = n_j$; when unequal, a harmonic mean is often used). The statistic q is compared to critical values from the Studentized range distribution.

Tukey's HSD.

Tukey's Honest Significant Difference method controls the family-wise error rate at α when making all pairwise comparisons of group means. It computes a confidence interval for each pairwise difference:

$$(\bar{x}_j - \bar{x}_{j'}) \pm q^* \cdot \sqrt{\frac{MSE}{n^*}}$$

where q^* is the critical value from the Studentized range distribution with k groups and $n-k$ degrees of freedom. If the confidence interval does not contain zero, the difference is discernible.

18.3.2 Example: MLB batting positions

After the ANOVA for MLB batting data ($F = 5.077, p = 0.0066$), we rejected H_0 and want to identify which position pairs differ in on-base percentage. Tukey's HSD produces the following pairwise comparisons:

Comparison	Difference	95% CI	Discernible?
IF - OF	-0.001	(-0.011, 0.008)	No
C - OF	-0.018	(-0.032, -0.004)	Yes
C - IF	-0.016	(-0.030, -0.003)	Yes

Catchers have a discernibly lower mean on-base percentage than both outfielders and infielders: neither the C − OF interval nor the C − IF interval contains zero, even after Tukey's adjustment for making three comparisons. The difference between infielders and outfielders is not discernible – its interval comfortably includes zero. So the discernible ANOVA result is driven by catchers, whose average OBP is roughly 0.016 to 0.018 lower than that of the other two positions.

18.4 Interpreting post-hoc results

Post-hoc tests (like Bonferroni or Tukey's HSD) are typically conducted *after* a discernible ANOVA result. Several important principles guide their interpretation:

18.4.1 When to use post-hoc tests

A common convention is to conduct post-hoc pairwise comparisons only after ANOVA rejects the null hypothesis. If the ANOVA p-value is not discernible, many analysts treat hunting through pairwise differences as a form of data snooping. This is a convention rather than a strict requirement: Tukey's HSD controls the family-wise error rate on its own, whether or not an ANOVA is run first.

18.4.2 Reporting results

When reporting results of multiple comparisons, it is common to use **compact letter displays**. Groups that share a letter are not discernibly different from each other:

Group	Mean	Grouping
A	75.1	ab
B	72.0	a
C	78.9	b

In this notation, Groups A and B share the letter "a", so they are not discernibly different, and Groups A and C share the letter "b", so they are not discernibly different either. Groups B and C share no letter, so they are discernibly different. Group A, sharing a letter with both B and C, is not discernibly different from either.

18.4.3 Confidence intervals for differences

An especially useful way to present multiple comparison results is through confidence intervals for pairwise differences. If a confidence interval contains zero, the two groups are not discernibly different. If the interval is entirely positive or entirely negative, the groups differ.

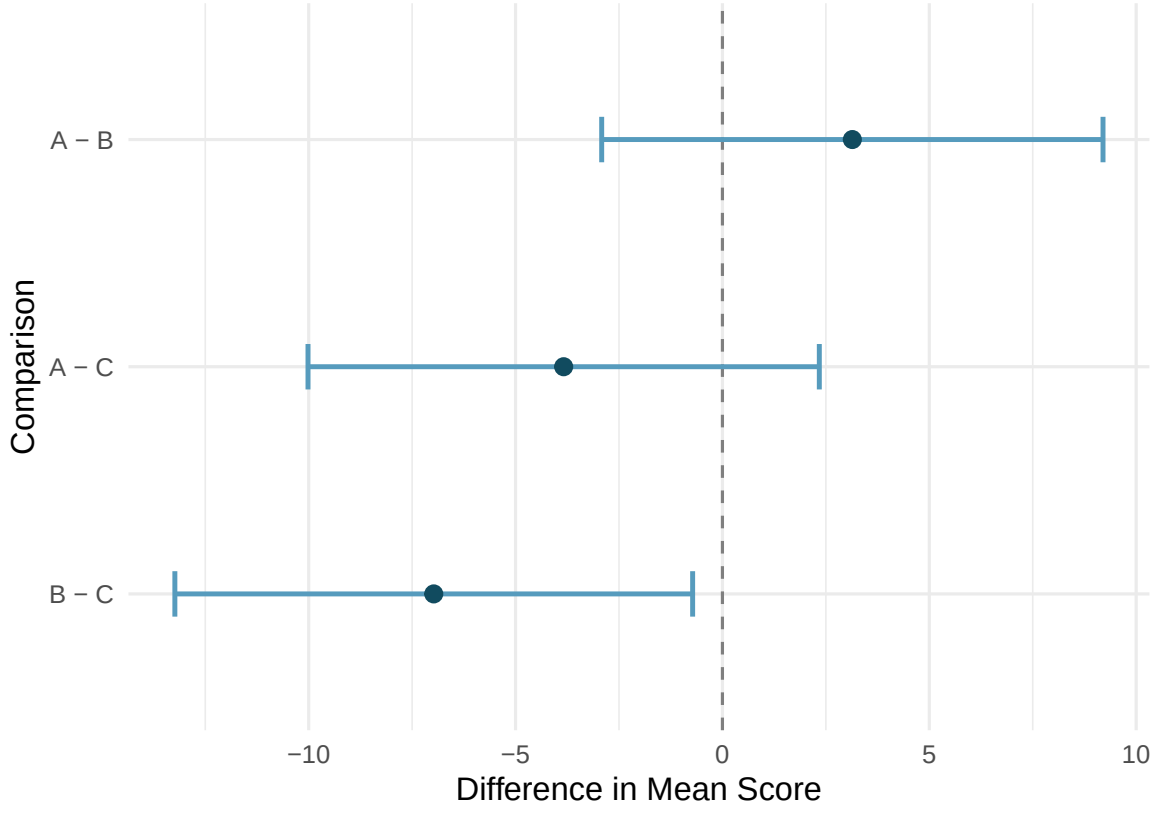


Figure 18.1: Tukey HSD 95% confidence intervals for pairwise differences in mean exam scores. The interval for B vs. C does not contain zero, indicating a discernible difference. ↗ Try this live — see every pair's Tukey/Bonferroni CI in a live forest plot.

18.4.4 Common mistakes

1. **Running multiple pairwise tests without correction.** This inflates the Type I error rate. Always use a correction method when conducting multiple comparisons.
2. **Hunting for pairwise differences after a non-discernible ANOVA.** By common convention, if the overall F-test does not reject H_0, we do not go looking for pairwise differences. If you do examine them, use a method such as Tukey's HSD that controls the family-wise error rate, and report the results as exploratory.
3. **Ignoring practical importance.** A statistically discernible difference may be too small to matter in practice. Always consider the magnitude of the differences, not just the p-values.
4. **Over-interpreting non-discernible pairwise comparisons.** A non-discernible pairwise comparison after a discernible ANOVA does not mean the groups are equal – it means we do not have enough evidence to declare them different individually.

18.5 Bonferroni vs. Tukey: When to use which

Both methods control the family-wise error rate, but they have different strengths:

Feature	Bonferroni	Tukey's HSD
Generality	Any set of tests	Pairwise comparisons of means only
Power	More conservative (less powerful)	Less conservative (more powerful)
Assumptions	None beyond the individual tests	Equal group variances (same as ANOVA)
Best for	Small number of planned comparisons	All pairwise comparisons after ANOVA

Rule of thumb: If you are comparing all pairs of group means after a discernible ANOVA, use Tukey's HSD. If you have a small number of pre-planned comparisons (not necessarily all pairs), the Bonferroni correction is appropriate.

18.6 Chapter review

18.6.1 Summary

When ANOVA rejects the null hypothesis, the natural follow-up question is: which groups differ? Performing many pairwise tests without correction inflates the family-wise error rate, making false positives likely. The Bonferroni correction divides the discernibility level by the number of comparisons, which is simple but conservative. Tukey's HSD method is specifically designed for all pairwise comparisons of means after ANOVA and is typically more powerful. Both methods control the family-wise error rate at the desired level.

18.6.2 Key terms

- Multiple comparisons problem
- Family-wise error rate (FWER)
- Bonferroni correction

- Tukey's Honest Significant Difference (HSD)
- Post-hoc test
- Compact letter display

18.7 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **True / False: ANOVA, I.** Determine if the following statements are true or false in ANOVA, and explain your reasoning for statements you identify as false.

a. As the number of groups increases, the modified discernibility level for pairwise tests increases as well.

b. As the total sample size increases, the degrees of freedom for the residuals increases as well.

c. The constant variance condition can be somewhat relaxed when the sample sizes are relatively consistent across groups.

d. The independence assumption can be relaxed when the total sample size is large.

2. **True / False: ANOVA, II.** Determine if the following statements are true or false, and explain your reasoning for statements you identify as false.

If the null hypothesis that the means of four groups are all the same is rejected using ANOVA at a 5% discernibility level, then...

a. we can then conclude that all the means are different from one another.

b. the standardized variability between groups is higher than the standardized variability within groups.

c. the pairwise analysis will identify at least one pair of means that are discernibly different.

d. the appropriate α to be used in pairwise comparisons is $0.05 \ / \ 4 = 0.0125$ since there are four groups.

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Frame the problem.** A researcher measures reaction time under **five** different lighting conditions and wants to know which conditions produce different mean reaction times. She plans to run all $\binom{5}{2} = 10$ pairwise t-tests, each at $\alpha = 0.05$.
   a. What is the **per-test** Type I error rate? What is the **family-wise** Type I error rate, and how do they differ?
   b. Under the assumption that all 10 null hypotheses are true and the tests are independent, what is the probability of at least one false positive across the 10 tests?
   c. A colleague suggests: "Just report the per-test α; don't worry about the family-wise rate." Why is that unacceptable when the goal is to identify *which* conditions differ?

2. **Choose the procedure.** For each scenario, name the appropriate approach — **Tukey HSD**, **Bonferroni**, **just the ANOVA F-test (no post-hoc needed)**, or **not a multiple-comparisons problem**.
   a. ANOVA rejects H_0 across 4 fertilizer brands; you want to identify which brands differ from which.
   b. ANOVA fails to reject H_0 across 4 fertilizer brands.
   c. You have 3 groups but pre-specified only *two* comparisons of interest before seeing the data (e.g., new-treatment vs. control, and new-treatment vs. standard).
   d. You want to compare mean scores across two teaching methods.
   e. You want to test the same difference-in-means for 100 different genes to see which are associated with a disease.

3. **Check conditions before you trust the post-hoc.** Open the Multiple Comparisons tool loaded with the classic `insectsprays` data (6 insecticides, count of insects found on each treated plot).
   a. Confirm from the tool's display: how many groups? What is the smallest per-group n? Is that enough for the pairwise procedure to work?
   b. What conditions does the Tukey HSD test require, beyond just "3+ groups"?
   c. The tool's output shows the ANOVA table above the pairwise panel. Why is it appropriate to *inspect* the omnibus F-test first, even if you plan to do pairwise comparisons?

4. **Run it and interpret it.** Stay in the Multiple Comparisons tool. Make sure `?posthoc=tukey` is active.
   a. Report the omnibus F statistic and p-value. Does the ANOVA reject H_0?
   b. In the pairwise table, count how many pairs have **adjusted p-values below 0.05**. How many pairs are declared *not* discernibly different?
   c. Read the forest plot: which pairs' CIs cross zero, and which do not? Group the 6 sprays into "equivalence classes" (letters that don't discernibly differ from each other but do differ from other classes).
   d. A student writes: "Sprays A and B differ from Spray C at family-wise $\alpha = 0.05$." Write the *correct* sentence including which quantity has family-wise coverage, and what "family-wise" actually means for the interpretation.
   e. The tool warns *not* to run post-hoc if the omnibus F is not significant. Why?

5. **Bonferroni vs. Tukey — when do they agree?** Switch the tool between `?posthoc=tukey` and `?posthoc=bonferroni` (or use the toggle).
   a. Compare the widths of the pairwise 95% CIs under Tukey vs. Bonferroni on the `insectsprays` data. Which method gives **narrower** intervals?

b. Repeat on a smaller dataset (e.g., 3 groups, small n). Do the two methods still favor Tukey by the same margin?
c. Why does Tukey give tighter intervals than Bonferroni **when the goal is all pairwise comparisons of means**? Tie your answer to what each method is *actually* controlling.
d. When would you prefer **Bonferroni** over Tukey?
e. A student is confused because on one dataset Bonferroni declares an extra pair discernibly different that Tukey does not. What could explain that, and which conclusion should you trust?

Chapter 19

Correlation

When we have two quantitative variables, we often want to understand the relationship between them. Does one variable increase as the other increases? Does it decrease? Or is there no apparent pattern? In this chapter, we introduce tools for visualizing and quantifying the strength and direction of the linear relationship between two quantitative variables. We begin with scatterplots, then introduce the correlation coefficient r, discuss its properties and limitations, and address the ever-important distinction between correlation and causation.

19.1 Scatterplots

A **scatterplot** displays the relationship between two quantitative variables. Each point represents a single observation, with one variable on the horizontal axis and the other on the vertical axis. Scatterplots are the most useful tool for visualizing the relationship between two quantitative variables.

When examining a scatterplot, we look for four key features:

1. **Direction**: Is the relationship positive (upward trend), negative (downward trend), or neither?
2. **Form**: Is the relationship linear (points cluster around a line) or nonlinear (points follow a curve)?
3. **Strength**: How tightly do the points cluster around the trend? A strong relationship has little scatter; a weak relationship has lots of scatter.
4. **Unusual observations**: Are there any outliers that deviate from the overall pattern?

Researchers captured 104 brushtail possums in Australia and took body measurements before releasing them. Consider the relationship between total body length (cm) and head length (mm).

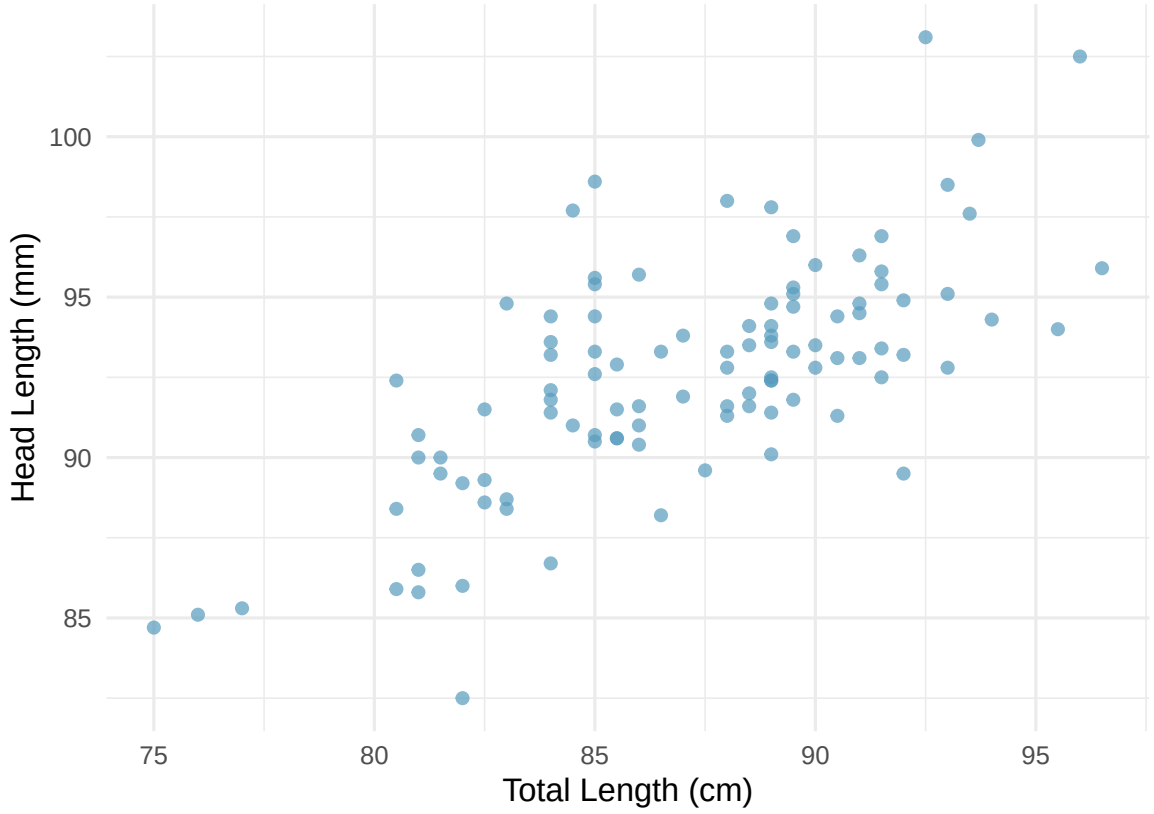


Figure 19.1: A scatterplot showing head length against total length for 104 brushtail possums. The relationship is moderately strong, positive, and linear. Open in StatLens

Describe the relationship between total length and head length.

The relationship is **positive** (possums with longer total lengths tend to have longer heads), **linear** (the points cluster around a straight line, not a curve), and **moderate in strength** (the points are somewhat scattered around the trend, but the pattern is clear). There do not appear to be any extreme outliers.

19.1.1 When linear models are not appropriate

There are cases where fitting a straight line to the data is not helpful, even if there is a clear relationship between the variables.

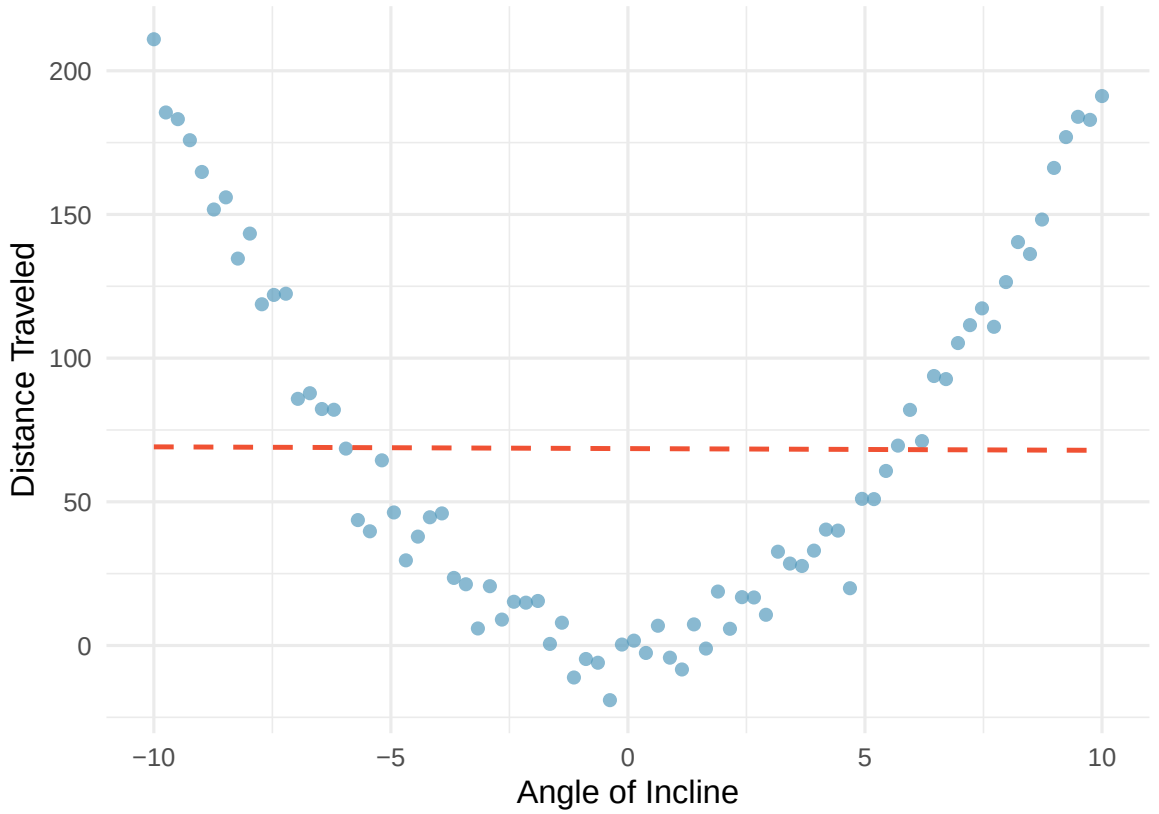


Figure 19.2: A scatterplot showing a clear quadratic (U-shaped) relationship between angle of incline and distance traveled. The best-fitting straight line is flat and does not capture the pattern at all.

When the relationship is nonlinear, fitting a straight line can be misleading. We must always examine scatterplots before computing correlations or fitting linear models.

19.2 The correlation coefficient

We need a way to quantify the strength and direction of a linear relationship. The **correlation coefficient** r does exactly this.

Correlation: strength of a linear relationship.

Correlation, denoted r, always takes values between -1 and 1 and describes the strength and direction of the linear relationship between two variables.

The correlation is unitless and is not affected by linear changes in the units of measurement (e.g., converting from inches to centimeters).

The formula for the correlation coefficient is:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s_x}\cdot\frac{y_i - \bar{y}}{s_y}$$

where $\bar{x}$ and $\bar{y}$ are the sample means, and s_x and s_y are the sample standard deviations for each variable. In practice, we always use a computer or calculator to compute r.

19.2.1 Interpreting correlation values

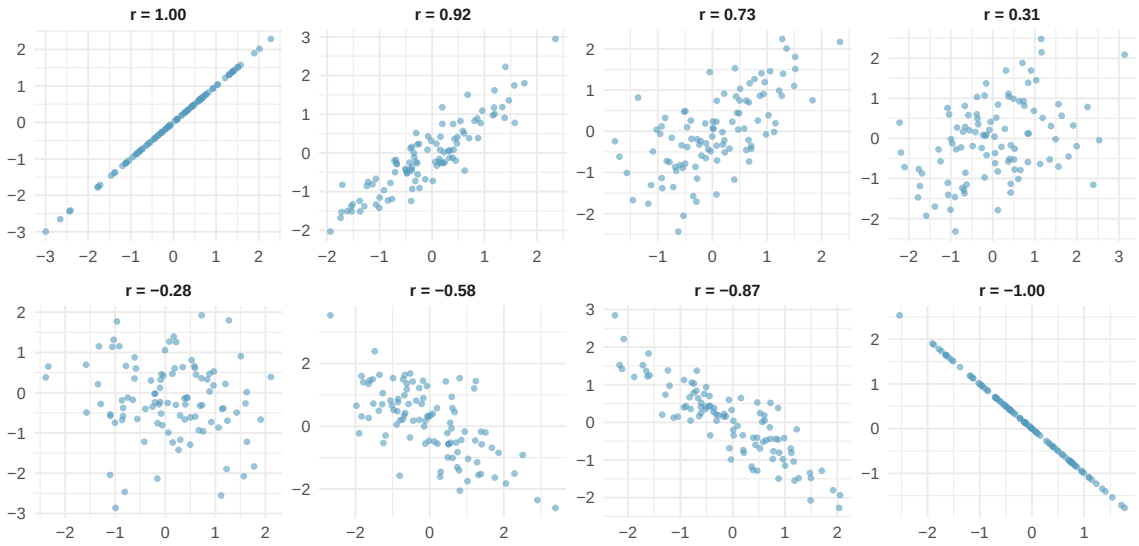


Figure 19.3: Eight scatterplots showing different correlation values. The first row shows positive relationships with r values of 1.00, 0.92, 0.73, and 0.31. The second row shows negative relationships with r values of -0.28, -0.58, -0.87, and -1.00. ↗ Try this live — play Guess the Correlation — build your feel for how r maps to scatter.

Key patterns to notice:

- When $r = 1$ or $r = -1$, all points fall exactly on a straight line (a perfect linear relationship).
- When r is close to $+1$ or -1, the relationship is strong and linear.
- When r is close to 0, there is no linear relationship (but there could still be a nonlinear relationship!).
- Positive r means both variables tend to increase together; negative r means one tends to decrease as the other increases.

19.2.2 Properties of correlation

The correlation coefficient has several important properties:

1. **Range:** $-1 \leq r \leq 1$.
2. **Unitless:** Correlation has no units. Changing the units of x or y (e.g., from cm to inches) does not change r.
3. **Symmetric:** The correlation of x with y is the same as the correlation of y with x.
4. **Not affected by linear transformations:** Adding a constant to or multiplying a variable by a positive constant does not change r.
5. **Only measures linear relationships:** Correlation can be misleading for nonlinear relationships.

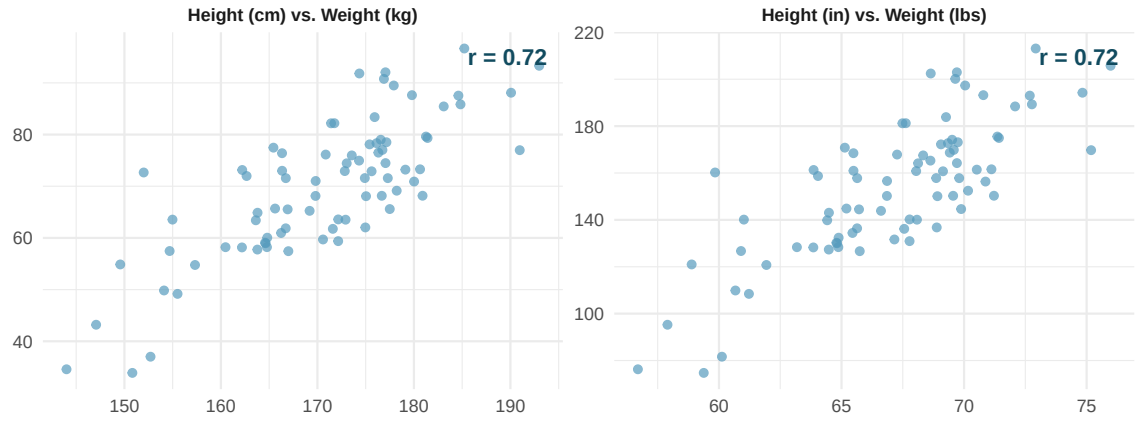


Figure 19.4: Two scatterplots showing the same relationship between weight and height, measured in different units (kg/cm vs. lbs/in). The correlation r = 0.72 is the same in both plots.

19.2.3 Correlation does not capture nonlinear relationships

The correlation coefficient is designed to measure the strength of *linear* trends. Nonlinear trends, even when strong, can produce misleading correlations.

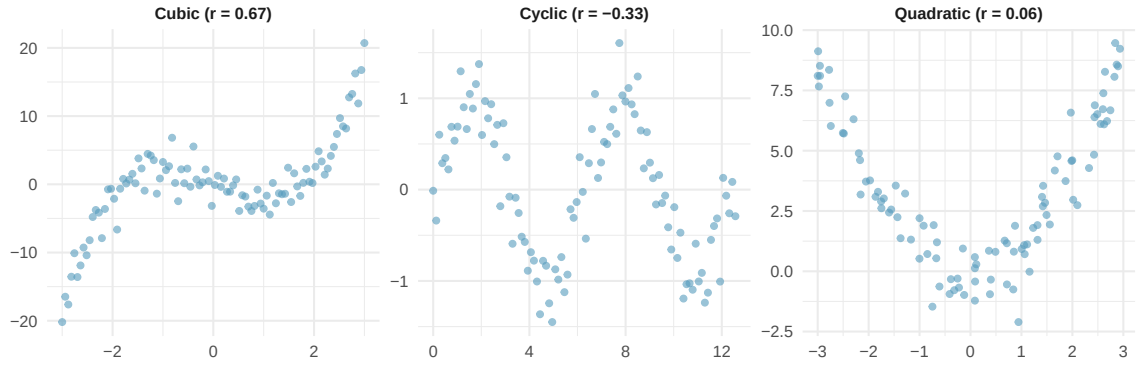


Figure 19.5: Three scatterplots showing strong nonlinear relationships (quadratic, cyclic, and a complex curve) that have weak correlations.

No straight line is a good fit for any of the datasets in the figure above. If you drew nonlinear curves on each plot, what general features would be important in your fit?

Show answer

The curves should be close to most points and reflect the overall nonlinear trend in the data.

Correlation only measures linear relationships. Always examine a scatterplot before computing or interpreting a correlation. A correlation near zero does not necessarily mean there is no relationship – it means there is no *linear* relationship. There may still be a strong nonlinear association.

Examine the scatterplots below showing the relationships between various crop yields across countries. Each point is a country, and each variable is the percentage of that country's total yield over the

past 50 years that came from the given crop (countries that did not grow a crop are omitted from that plot). Order the plots from strongest negative to strongest positive linear relationship.

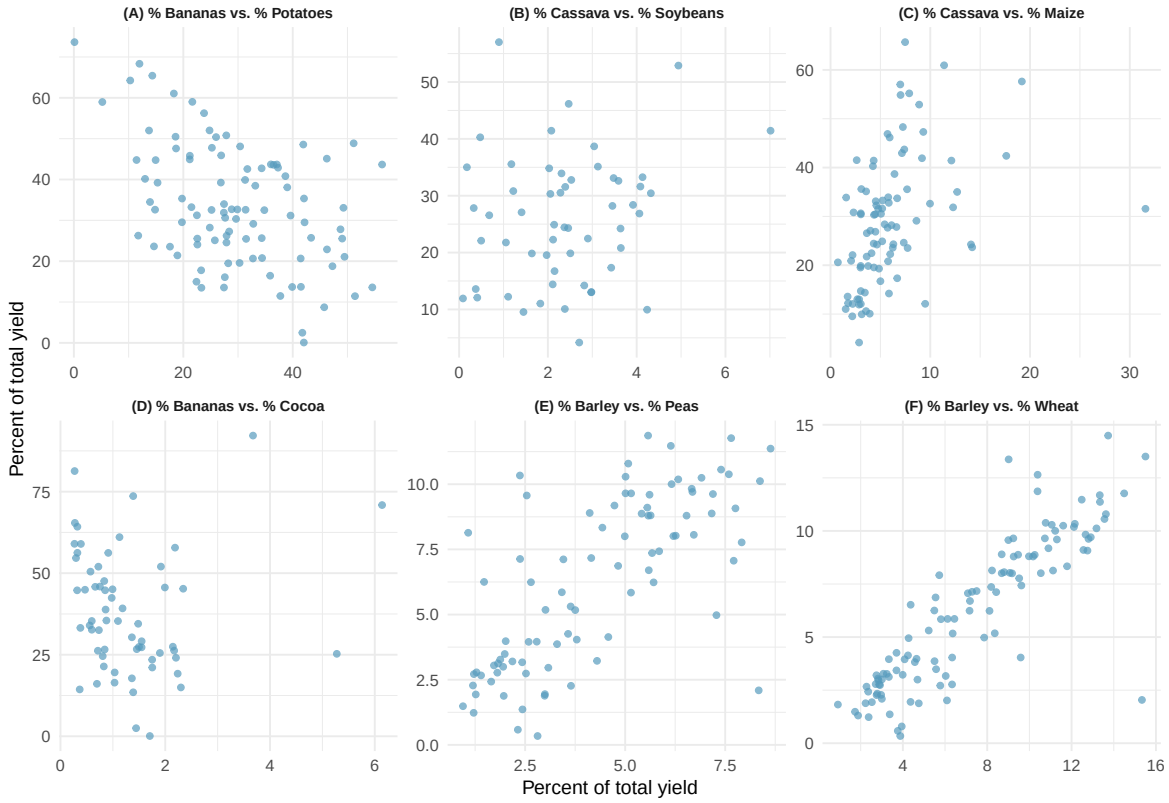


Figure 19.6: Six scatterplots of crop yield percentages across countries. Each shows a different pair of crop types with varying degrees of correlation.

Ordering from most negative to most positive correlation: the plots showing (A) bananas vs. potatoes (r = -0.45), (D) bananas vs. cocoa ($r = 0.01$), (B) cassava vs. soybeans ($r = 0.19$), (C) cassava vs. maize ($r = 0.41$), (E) barley vs. peas ($r = 0.71$), and (F) barley vs. wheat ($r = 0.86$).

19.3 Correlation and causation

Correlation does not imply causation.

Just because two variables are correlated does not mean that one causes the other. There may be a **confounding variable** (also called a lurking variable) that is associated with both x and y, creating a correlation between them even though there is no direct causal link.

Studies have found a positive correlation between the number of firefighters sent to a fire and the amount of damage the fire causes. Does this mean sending more firefighters causes more damage?

No. The confounding variable is the **severity of the fire**. Larger fires cause more damage *and* require more firefighters. The number of firefighters does not cause the damage; both are driven by the severity of the fire.

To establish causation, we need one of the following:

1. A **randomized experiment** where the explanatory variable is randomly assigned.
2. Strong theoretical justification combined with careful control of confounding variables.

Observational studies can reveal associations (correlations), but they cannot establish causation on their own.

19.4 Sampling variability of correlation

Like any sample statistic, the correlation coefficient r computed from a sample is just an estimate of the true population correlation ρ (the Greek letter "rho"). Different samples from the same population will produce different values of r.

When the sample size is small, r can vary substantially from sample to sample, even when the true population correlation is zero. This means we should be cautious about interpreting correlations from small samples.

Suppose the true population correlation between two variables is $\rho = 0$ (no linear relationship). If we draw random samples of size $n = 10$, the sample correlation r could reasonably fall anywhere between about -0.63 and $+0.63$ just by chance.

With larger samples ($n = 100$), the sampling variability shrinks, and r would typically fall between about -0.20 and $+0.20$.

This illustrates why a formal test of discernibility is important: we need to determine whether an observed correlation is large enough to be unlikely under the null hypothesis that $\rho = 0$.

The next chapter, on linear regression, runs a randomization test of whether the population slope of a regression line is zero. That test is equivalent to testing whether the population correlation is zero, so it answers the same question about ρ.

See it in action. Guess the Correlation — a short scored warmup that calibrates your eye on what "$r = 0.7$" actually *looks* like. For inference, the Randomization Test for Correlation shuffles y to break the x–y pairing and builds a null distribution of r. For general exploration, use the Regression Explorer.

19.5 Chapter review

19.5.1 Summary

In this chapter, we introduced scatterplots as the primary tool for visualizing the relationship between two quantitative variables. We then defined the correlation coefficient r as a measure of the strength and direction of the linear relationship. Key properties of r include: it ranges from -1 to $+1$, it is unitless, it is symmetric, and it measures only linear relationships. We emphasized that correlation does not imply causation – confounding variables can create spurious associations. Finally, we noted that the sample correlation is subject to sampling variability, especially in small samples.

19.5.2 Key terms

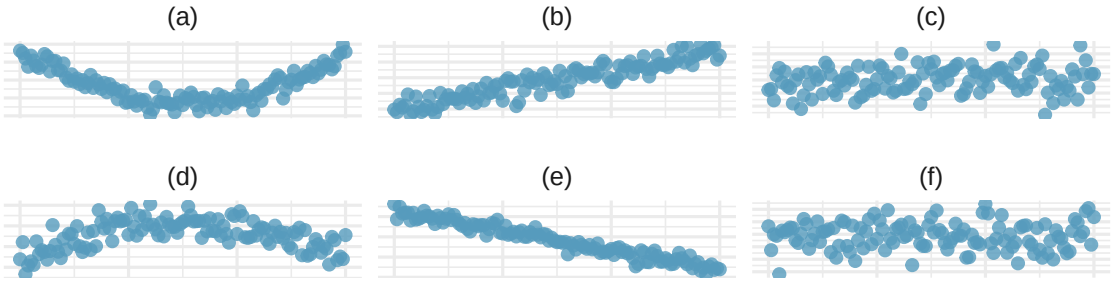
- Scatterplot
- Correlation (r)
- Positive association / negative association
- Nonlinear relationship
- Confounding variable

- Correlation does not imply causation
- Population correlation (ρ)

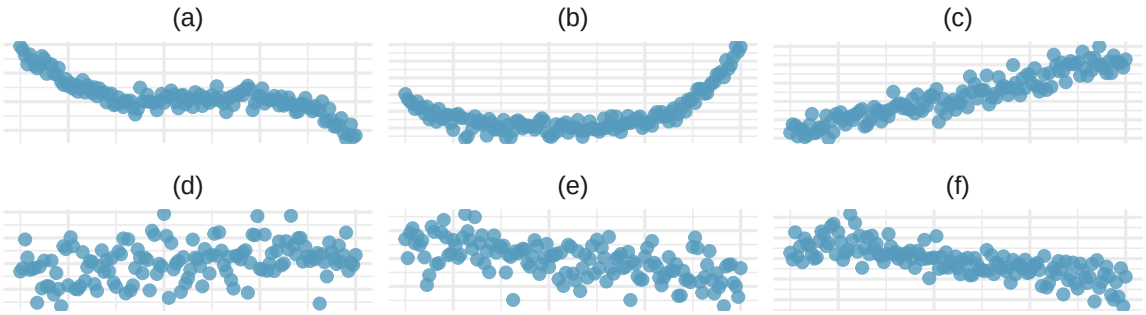
19.6 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

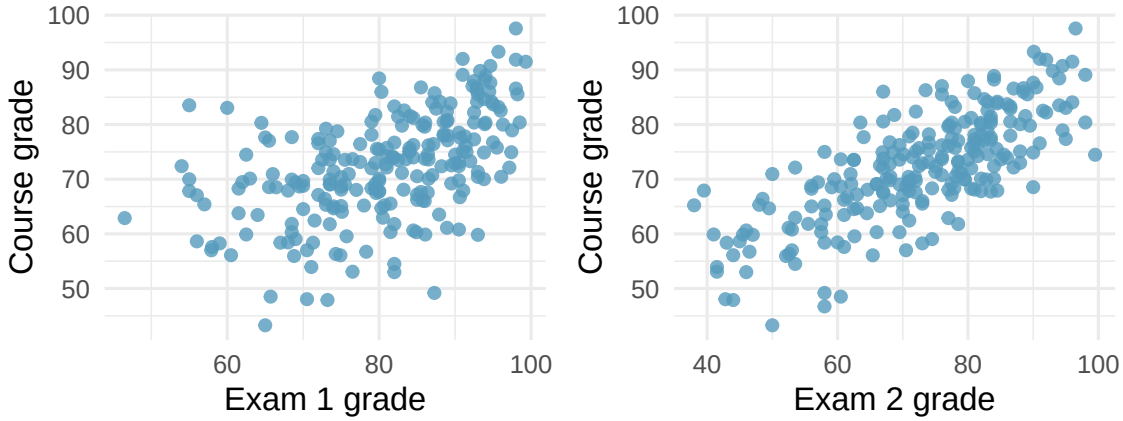
1. **Identify relationships, I.** For each of the six plots, identify the strength of the relationship (e.g., weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.



2. **Identify relationships, II.** For each of the six plots, identify the strength of the relationship (e.g., weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.



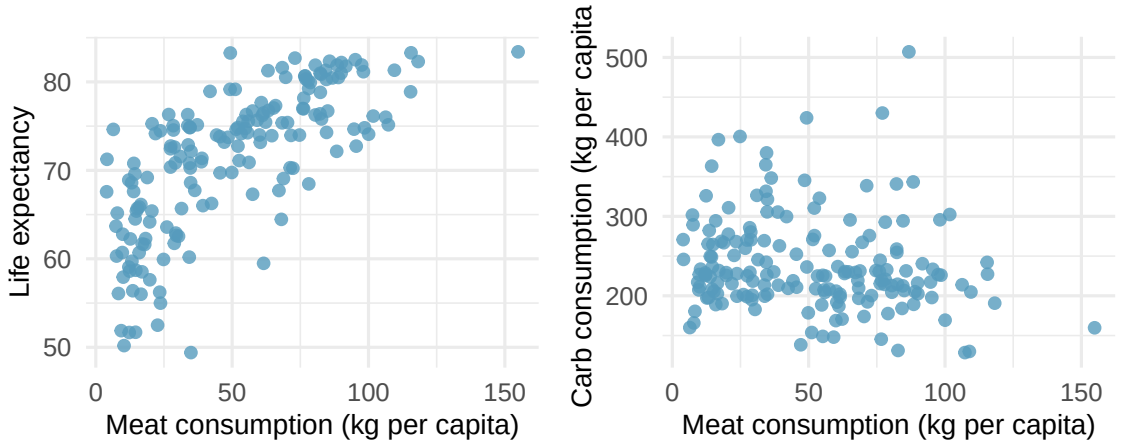
3. **Midterms and final.** The two scatterplots below show the relationship between the overall course average and two midterm exams (Exam 1 and Exam 2) recorded for 233 students during several years for a statistics course at a university.



a. Based on these graphs, which of the two exams has the strongest correlation with the course grade? Explain.

b. Can you think of a reason why the correlation between the exam you chose in part (a) and the course grade is higher?

4. **Meat consumption and life expectancy.** In data collected for You et al. (2022), total meat intake is associated with life expectancy (at birth) in 175 countries. Meat intake is measured in kg per capita per year (averaged over 2011 to 2013). Additionally, the authors collected data on carbohydrate crops (e.g., cereals, root, sugar, etc.) in kg per capita per year (averaged over 2011 and 2013). The scatterplot on the left shows the life expectancy at birth plotted against the per capita meat consumption. The scatterplot on the right shows the amount of carbohydrate consumption plotted against the per capita meat consumption.



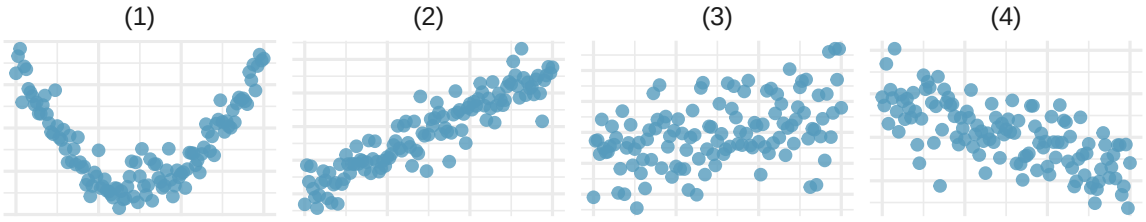
a. Describe the relationship between meat consumption and life expectancy.

b. Describe the relationship between meat consumption and carbohydrate consumption.

c. Which plot shows a stronger correlation? Explain your reasoning.

d. Data on consumption were originally collected in kilograms. How would the plots above and the correlation values change if the consumption variables were converted to pounds?

5. **Match the correlation, I.** Match each correlation to the corresponding scatterplot.



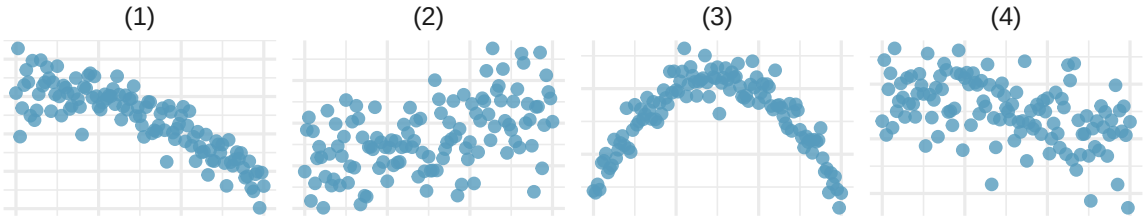
a. $r = -0.7$

b. $r = 0.45$

c. $r = 0.06$

d. $r = 0.92$

6. **Match the correlation, II.** Match each correlation to the corresponding scatterplot.



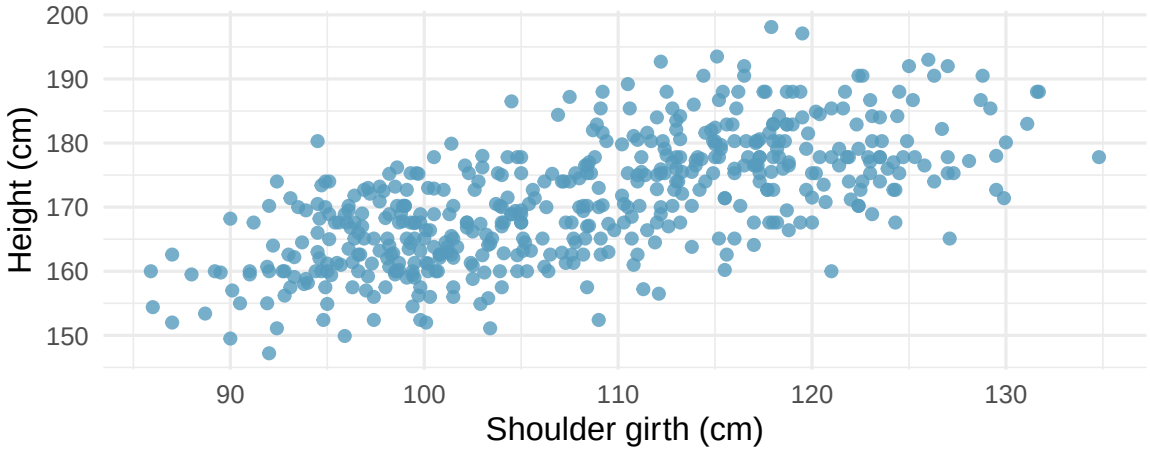
a. $r = 0.49$

b. $r = -0.48$

c. $r = -0.03$

d. $r = -0.85$

7. **Body measurements, correlation.** Researchers studying anthropometry collected body and skeletal diameter measurements, as well as age, weight, height and sex for 507 physically active individuals. The scatterplot below shows the relationship between height and shoulder girth (circumference of shoulders measured over deltoid muscles), both measured in centimeters. (Heinz et al. 2003)

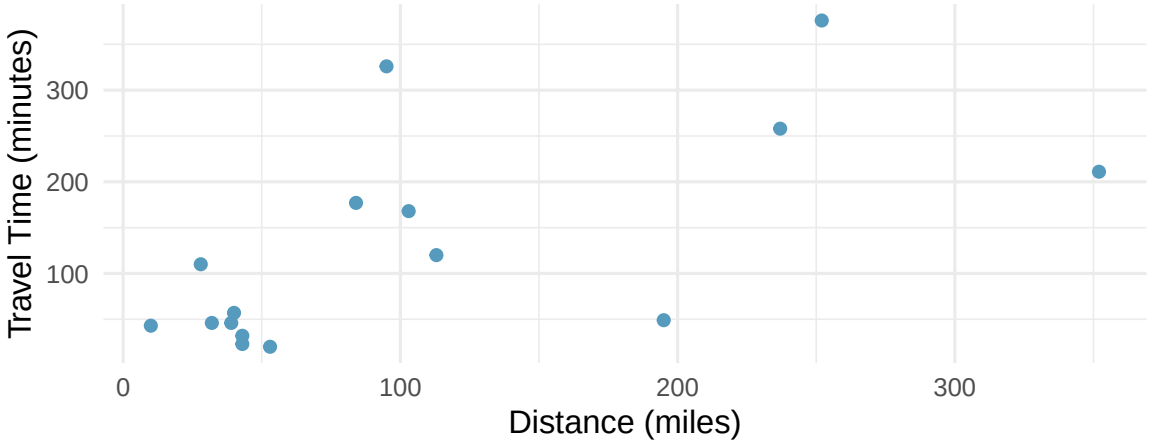


a. Describe the relationship between shoulder girth and height.

b. How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?

8. **Compare correlations.** Eduardo and Rosie are both collecting data on number of rainy days in a year and the total rainfall for the year. Eduardo records rainfall in inches and Rosie in centimeters. How will their correlation coefficients compare?

9. **The Coast Starlight, correlation.** The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes).

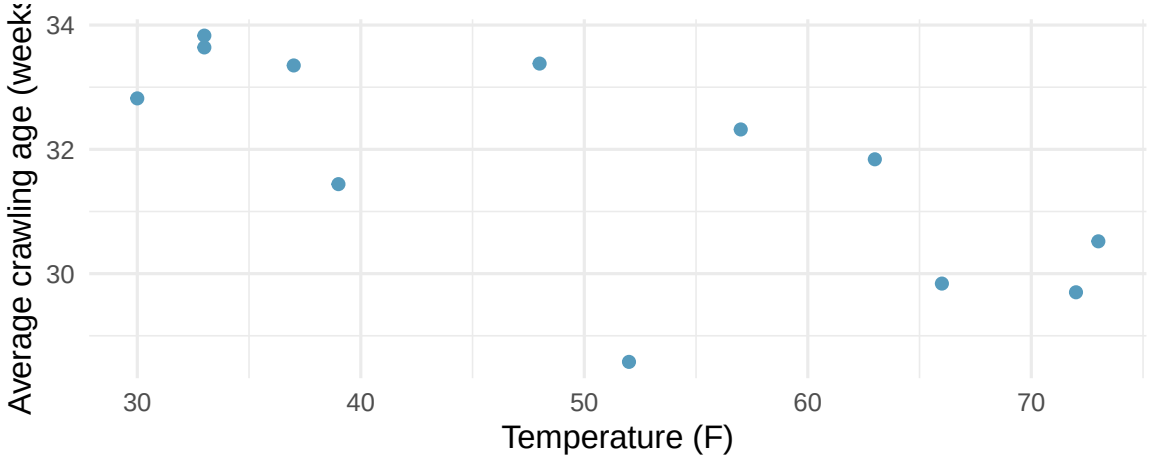


a. Describe the relationship between distance and travel time.

b. How would the relationship change if travel time was instead measured in hours, and distance was instead measured in kilometers?

c. Correlation between travel time (in miles) and distance (in minutes) is $r = 0.636$. What is the correlation between travel time (in kilometers) and distance (in hours)?

10. **Crawling babies, correlation.** A study conducted at the University of Denver investigated whether babies take longer to learn to crawl in cold months, when they are often bundled in clothes that restrict their movement, than in warmer months. Infants born during the study year were split into twelve groups, one for each birth month. We consider the average crawling age of babies in each group against the average temperature when the babies are six months old (that's when babies often begin trying to crawl). Temperature is measured in degrees Fahrenheit (F) and age is measured in weeks. (Benson 1993)



a. Describe the relationship between temperature and crawling age.

b. How would the relationship change if temperature was measured in degrees Celsius (C) and age was measured in months?

c. The correlation between temperature in F and age in weeks was $r = -0.70$. If we converted the temperature to C and age to months, what would the correlation be?

11. **Meat and carbohydrate consumption.** What would be the correlation between the per capita meat consumption and per capita carbohydrate consumption if, for each country, people always consumed

12. 3 kg more of meat than of carbohydrates each year?

13. 2 kg less of meat than of carbohydrates each year?

14. half as much meat as carbohydrates each year?

12. **Graduate degrees and salaries.** What would be the correlation between the annual salaries of people with and without a graduate degree at a company if, for a certain type of position, someone with a graduate degree always made

a. $5,000 more than those without a graduate degree?

b. 25% more than those without a graduate degree?

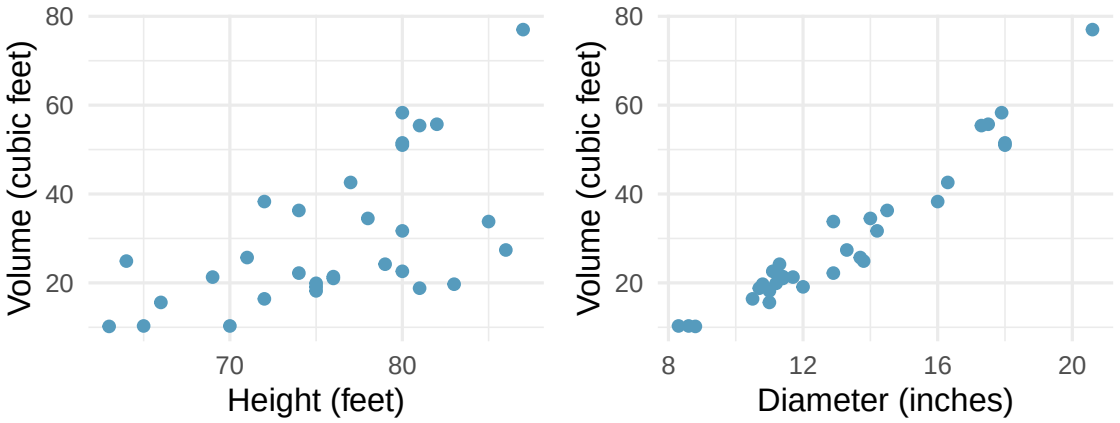
c. 15% less than those without a graduate degree?

13. **True / False.** Determine if the following statements are true or false. If false, explain why.

a. A correlation coefficient of -0.90 indicates a stronger linear relationship than a correlation of 0.5.

b. Correlation is a measure of the association between any two variables.

14. **Cherry trees.** The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground.

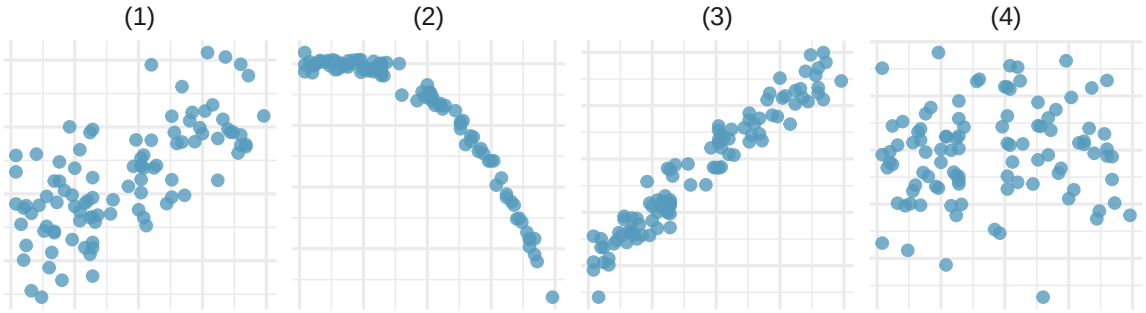


a. Describe the relationship between volume and height of these trees.

b. Describe the relationship between volume and diameter of these trees.

c. Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.

15. **Match the correlation, III.** Match each correlation to the corresponding scatterplot.



a. r = 0.69

b. r = 0.09

c. r = -0.91

d. r = 0.97

Datasets: `exam_grades` (openintro) | `corr_match` (openintro) | `bdims` (openintro) | `coast_starlight` (openintro) | `babies_crawl` (openintro) | `trees` (datasets)

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Calibrate your eye.** Open Guess the Correlation and play one round of at least 10 scatterplots.
   a. What is your score? How well did you do on the *strong* correlations ($|r| > 0.7$) versus the *weak* ones ($|r| < 0.3$)?
   b. Were you more often *over-* or *under*-estimating $|r|$? Why might that be? (Hint: most people see a slight pattern and call it $r = 0.6$ when it's actually around $r = 0.3$.)
   c. Two scatterplots can have the same r but tell very different stories — e.g., a clean line and a curve might both register the same r. What does this say about *only* reporting r without also showing the scatterplot?

2. **Choose the procedure.** For each scenario, name the most appropriate procedure — **Pearson correlation r**, **randomization test for $\rho = 0$**, **two-sample t-test**, or **not a correlation problem**.
   a. Is there a linear association between hours studied and exam score?
   b. Test whether the population correlation between two variables differs from zero.
   c. Is the *mean* exam score different for students who study with music vs. without?
   d. Visualize whether one variable tends to go up as another goes up.
   e. Test whether the *median* income is higher in one region than another.

3. **Check the conditions before you trust r.** Open the Regression Explorer with any two-quantitative-variable dataset.
   a. What three features of the scatterplot would make you uneasy about reporting r?
   b. Suppose the scatter shows a strong **curved** pattern but r comes out near 0. What's the right report — "no association" or "no *linear* association"?
   c. Suppose two outliers in the upper-right corner inflate r from 0.1 (without them) to 0.6 (with them). What should you do — report both? Trust one? Remove the outliers and not mention it?

4. **Run the test for $\rho = 0$.** Open the Randomization Correlation tool with any two-quantitative dataset.
   a. Report the observed r and the simulation p-value (after shuffling y many times).
   b. Write a one-sentence conclusion **in context**, naming both variables.
   c. Why does shuffling **only** y (or equivalently shuffling only x, but not both) simulate the world where $\rho = 0$?
   d. Which statement is the correct interpretation of a p-value of 0.04?
      i. The probability that the variables are unrelated is 4%.
      ii. If the variables were unrelated, we'd see a correlation at least as extreme as ours about 4% of the time by chance.
      iii. The correlation has only a 4% chance of being a real effect.

5. **Correlation, causation, and confounders.** A study reports a strong positive correlation between **ice-cream sales** and **drowning deaths** in summer months across U.S. cities.
   a. Could ice-cream sales *cause* drowning deaths? Could drowning deaths *cause* ice-cream sales? Name a likely **confounder** that makes both variables move together.
   b. Suppose a new study controls for that confounder and finds that ice-cream sales and drowning deaths are *uncorrelated* once the confounder is held constant. What does the original r now mean?

c. In general, what kind of **study design** would let you draw a causal conclusion from a correlation? (Open the Regression Slope tool if it helps to think about regression and randomization together.)
d. A news article reports "Coffee Linked to Lower Risk of Death" from an *observational* study. List two specific reasons you should not take this as a causal claim.

Chapter 20

Linear Regression

Linear regression is a powerful statistical technique for describing the relationship between two quantitative variables. Many people have some familiarity with regression just from reading the news, where straight lines are overlaid on scatterplots. In this chapter, we build on the ideas of correlation from the previous chapter to develop the least squares regression line. We learn how to interpret the slope and intercept, use residuals to assess model fit, quantify the strength of the model with R^2, and perform inference on the slope to test whether a linear relationship exists in the population.

20.1 Fitting a line to data

Linear regression is the statistical method for fitting a line to data where the relationship between two variables, x and y, can be modeled by a straight line with some error:

$$y = b_0 + b_1 x + e$$

The values b_0 and b_1 represent the model's **intercept** and **slope**, respectively, and the error is represented by e. These values are calculated from the data – they are sample statistics. If the observed data are a random sample from a target population, these values are point estimates for the population parameters β_0 and β_1.

When we use x to predict y, we call x the **predictor** variable (or explanatory variable) and y the **outcome** variable (or response variable). We often write the model without the error term when focusing on the predicted average outcome:

$$\hat{y} = b_0 + b_1 x$$

The "hat" on y signifies that this is a predicted (estimated) value.

It is rare for all data to fall perfectly on a straight line. Instead, data more commonly appear as a cloud of points, where the trend may be strong or weak.

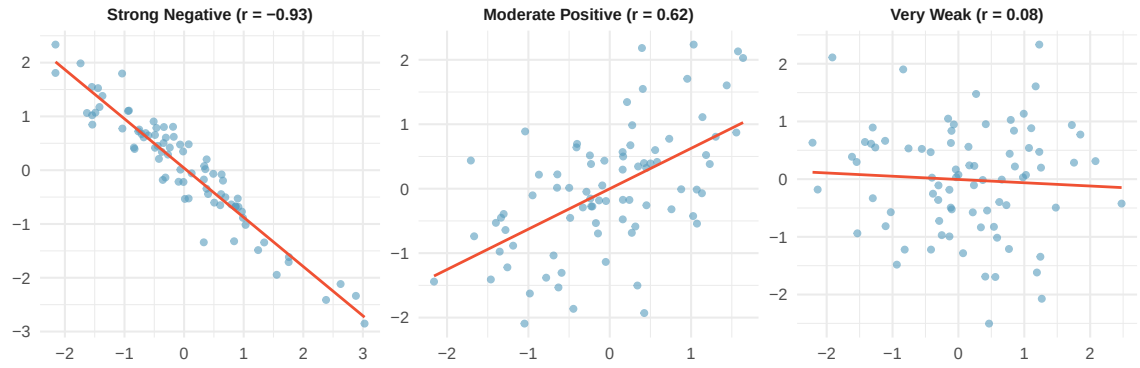


Figure 20.1: Three scatterplots with fabricated data showing a strong negative linear trend, a moderate positive linear trend, and a very weak trend.

There are also cases where fitting a straight line is inappropriate, even if there is a clear relationship between the variables. When the relationship is nonlinear, a linear model can be misleading.

20.1.1 Using linear regression to predict possum head lengths

Researchers captured 104 brushtail possums in Australia and took body measurements. We consider total body length (cm) as the predictor to predict head length (mm).

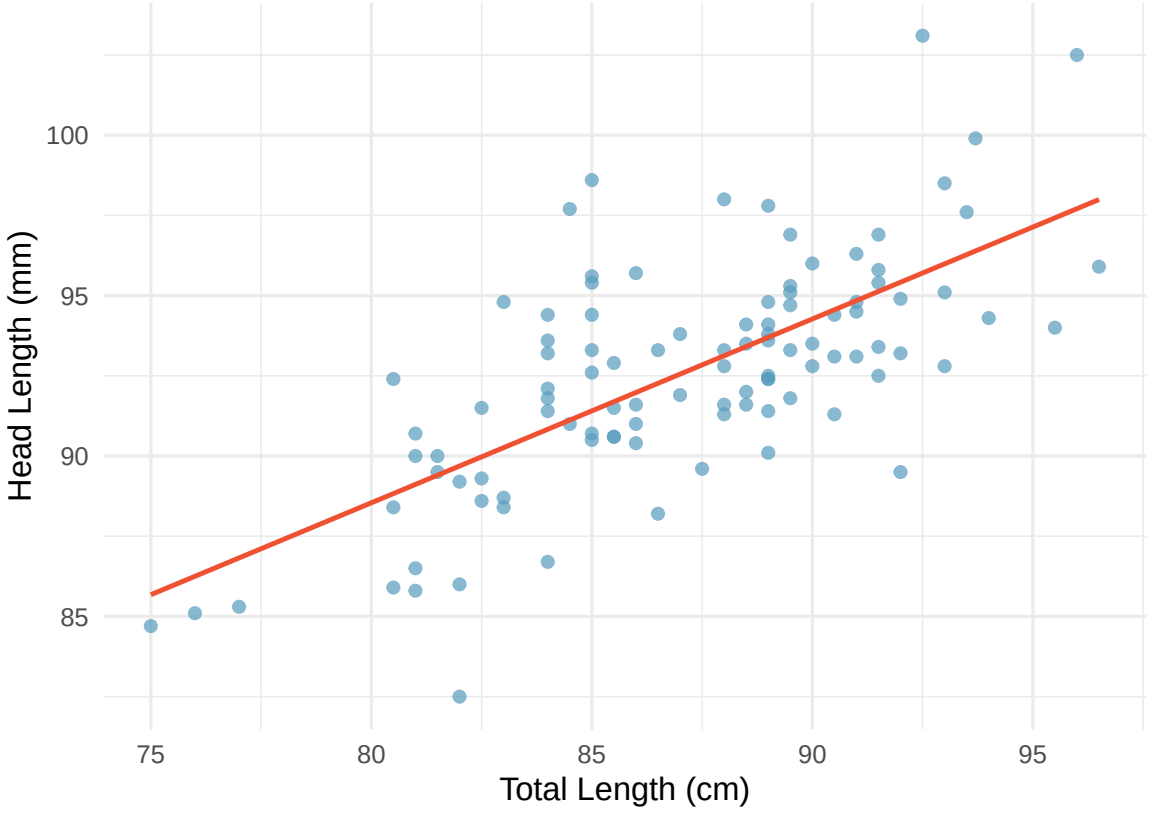


Figure 20.2: A scatterplot showing head length against total length for 104 brushtail possums, with a linear model fit to the data. Open in StatLens

The equation for this line is:

$$\hat{y} = 41 + 0.59x$$

We can use this line to make predictions. For instance, the equation predicts a possum with a total length of 80 cm will have a head length of:

$$\hat{y} = 41 + 0.59 \times 80 = 88.2 \text{ mm}$$

This estimate may be viewed as an average: the equation predicts that possums with a total length of 80 cm will have an average head length of 88.2 mm.

20.2 Residuals

Residuals are the leftover variation in the data after accounting for the model fit:

$$\text{Data} = \text{Fit} + \text{Residual}$$

Each observation has a residual. If an observation is above the regression line, its residual is positive. If it is below the line, the residual is negative.

Residual: Difference between observed and expected.

The residual of the ith observation (x_i, y_i) is the difference between the observed outcome and the predicted outcome:

$$e_i = y_i - \hat{y}_i$$

We identify $\hat{y}_i$ by plugging x_i into the model.

The linear fit is $\hat{y} = 41 + 0.59x$. Compute the residual for the observation $(76.0, 85.1)$.

First compute the predicted value: $\hat{y} = 41 + 0.59 \times 76.0 = 85.84$. The residual is $e = 85.1 - 85.84 = -0.74$ mm. The negative residual indicates that the model overpredicted head length for this possum.

Try this in StatLens →

If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates?

Show answer

If a model underestimates an observation, the residual is positive (the actual value exceeds the prediction). If it overestimates, the residual is negative.

20.2.1 Residual plots

Residuals are helpful in evaluating how well a linear model fits a dataset. We display them in a **residual plot**, where the horizontal axis shows the predicted values and the vertical axis shows the residuals.

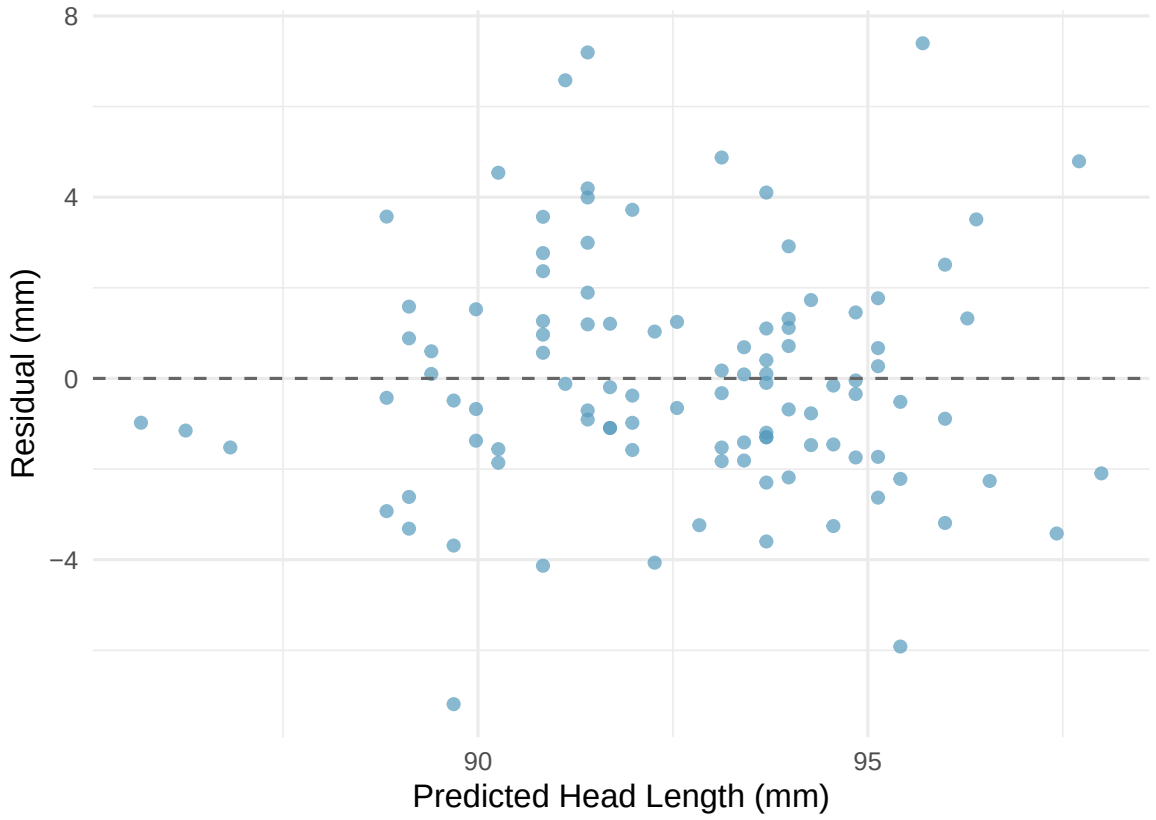


Figure 20.3: A residual plot for the possum model. The residuals appear randomly scattered around zero with no obvious pattern. Open in StatLens

What patterns should we look for in a residual plot?

- **No pattern (good):** Residuals scattered randomly around zero suggest the linear model is appropriate.
- **Curved pattern (bad):** A U-shape or other curve in the residuals indicates a nonlinear relationship that a straight line fails to capture.
- **Fan shape (bad):** Increasing or decreasing spread in the residuals indicates non-constant variability.

20.3 Least squares regression

Fitting linear models by eye is subjective. **Least squares regression** provides an objective, rigorous approach.

20.3.1 The objective: minimize squared residuals

We want a line with small residuals. The **least squares line** minimizes the sum of the squared residuals:

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

Why square the residuals rather than just sum their absolute values?

1. It is the most commonly used method.
2. It is widely supported in statistical software.
3. A residual twice as large is usually more than twice as bad. Squaring accounts for this.
4. The inference theory connecting the model to population parameters is most straightforward with least squares.

20.3.2 Interpreting the slope and intercept

For the Elmhurst College data, the least squares line predicting gift aid (in \$1,000s) from family income (in \$1,000s) is:

$$\widehat{\text{aid}} = 24.3 - 0.0431 \times \text{family\_income}$$

Term	Estimate	Std. Error	T statistic	p-value
(Intercept)	24.32	1.29	18.83	<0.0001
family_income	-0.0431	0.0108	-3.98	0.0002

What do the intercept and slope mean?

Slope ($b_1 = -0.0431$)**:** For each additional \$1,000 of family income, we would expect a student to receive \$43.10 less in gift aid, on average. The negative coefficient means higher income is associated with less aid. We must be cautious: this is an observational study, so we cannot interpret a causal connection.

Intercept ($b_0 = 24.32$)**:** The model predicts that a student whose family has no income would receive \$24,320 in gift aid, on average. This is meaningful because the family income for some students is near zero. In other applications, the intercept may have no practical meaning if $x = 0$ is outside the range of the data.

Interpreting parameters estimated by least squares.

The **slope** describes the estimated difference in the predicted average outcome of y if the predictor variable x were one unit larger.

The **intercept** describes the average outcome of y if $x = 0$, provided the linear model is valid at $x = 0$ (which may not be the case if $x = 0$ is far outside the observed data range).

20.3.3 Computing the slope and intercept from sample statistics

The slope can be computed using the correlation and the standard deviations of the two variables:

$$b_1 = \frac{s_y}{s_x} \cdot r$$

The least squares line always passes through the point $(\bar{x}, \bar{y})$. Using the point-slope form:

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

Identifying the least squares line from sample statistics.

- Estimate the slope: $b_1 = (s_y/s_x) \cdot r$.
- Find the intercept using the fact that $(\bar{x}, \bar{y})$ is on the line: $b_0 = \bar{y} - b_1\bar{x}$.

20.4 R-squared

We evaluated the strength of a linear relationship using r. More commonly, we describe the strength of a linear fit using R^2, called **R-squared** or the **coefficient of determination**.

Coefficient of determination: proportion of variability explained by the model.

R^2 measures the proportion of variation in the outcome variable y that is explained by the linear model with predictor x. Since r is always between -1 and 1, R^2 is always between 0 and 1.

$$R^2 = 1 - \frac{SSE}{SST}$$

where:

- $SST = \sum(y_i - \bar{y})^2$ is the **total sum of squares**, measuring total variability in y.
- $SSE = \sum(y_i - \hat{y}_i)^2 = \sum e_i^2$ is the **sum of squared errors**, measuring leftover variability after using the model.

For simple linear regression, $R^2 = r^2$.

For the Elmhurst data, the correlation is $r = -0.499$. What proportion of the variability in gift aid is explained by family income?

$R^2 = (-0.499)^2 = 0.249$, or about 25%. Family income explains about 25% of the variability in gift aid among these students.

If a linear model has a very strong negative relationship with $r = -0.97$, how much of the variation in the outcome is explained by the predictor?

Show answer

$R^2 = (-0.97)^2 = 0.94$, or about 94% of the variation is explained.

See it in action. Two guided walkthroughs: Fit a Line by Eye lab (OLS vs. LAD losses, noise-vs-uncertainty story) and Slope Inference walkthrough (LINE conditions → t-test → interpretation). Or open the Regression Explorer or Slope t-Test directly on your own data.

20.5 Inference for the slope

We have been computing sample statistics b_0 and b_1, but these are estimates of population parameters β_0 and β_1. We now turn to formal inference: testing whether a linear relationship exists in the population.

20.5.1 Hypotheses

- H_0: $\beta_1 = 0$. There is no linear relationship between x and y.
- H_A: $\beta_1 \neq 0$. There is a linear relationship between x and y.

20.5.2 Randomization test for the slope

If the null hypothesis is true ($\beta_1 = 0$), then x and y are not linearly related. We can simulate this by **permuting** the response variable: randomly shuffling the y values while keeping the x values fixed. This breaks any relationship between x and y.

To illustrate, we use the `births14` data from the **openintro** R package: a random sample of 1,000 US births from 2014, with each baby's birth weight (lbs) and length of gestation (weeks).

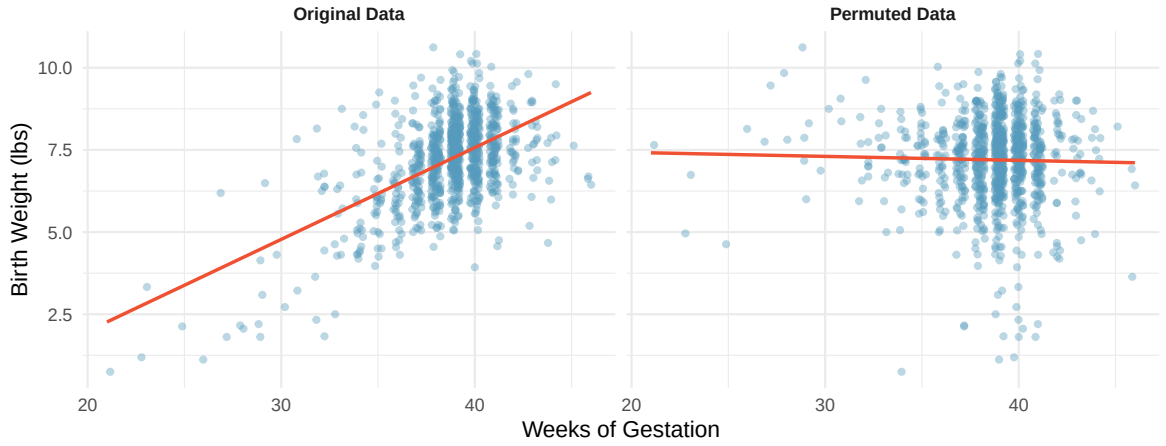


Figure 20.4: Original data for 1,000 US births in 2014, showing a positive linear relationship between weeks of gestation and birth weight, alongside the same data after permuting the weight variable, which destroys the relationship. Gestation is recorded in whole weeks, so the points are jittered horizontally to reduce overplotting. ↗ Try this live — permute the response yourself and compare to the observed pattern.

By repeating the permutation many times and computing the slope each time, we build a null distribution of slopes.

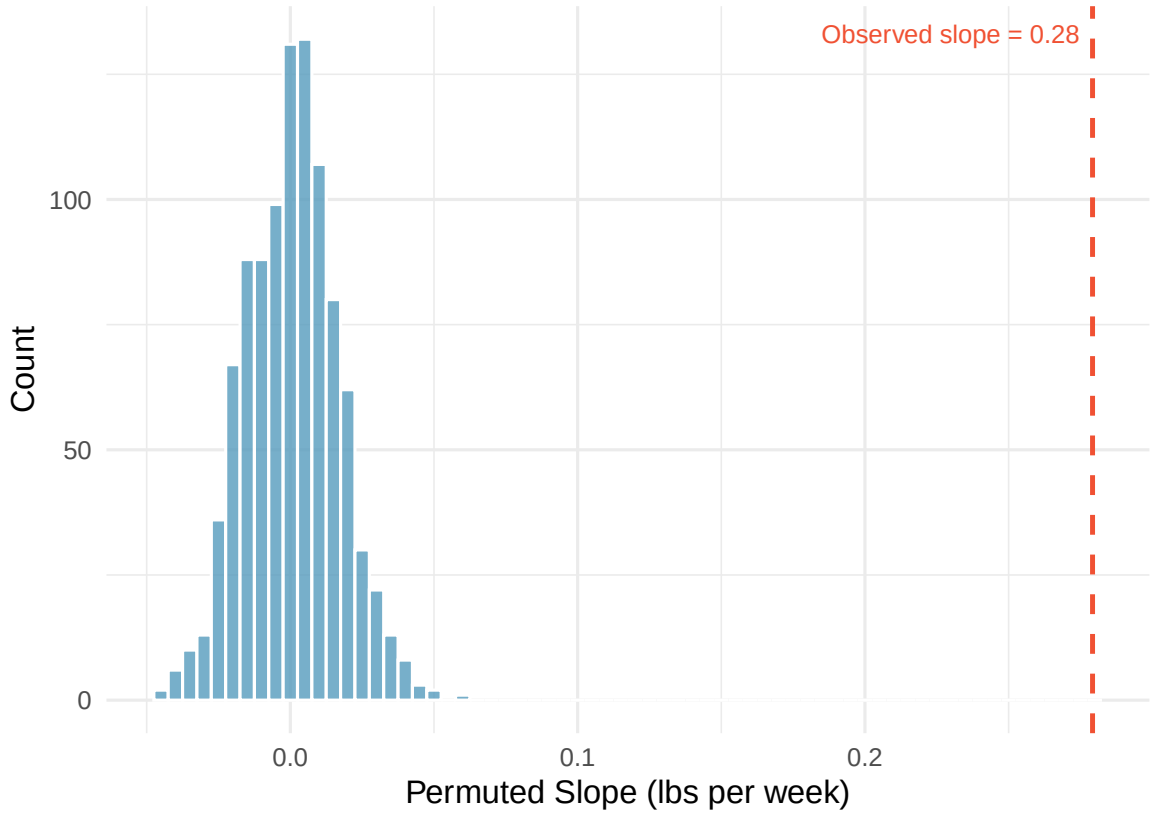


Figure 20.5: Histogram of slopes from 1,000 permutations of birth weight. The permuted slopes range from about −0.05 to +0.06 lbs per week, centered at zero. The observed slope of 0.28 lbs per week is far from this distribution. ↗ Try this live — build the null distribution live from permutations (the tool records the correlation r for each shuffle, which gives the same test as the slope).

The observed slope of 0.28 lbs per week is far from any value in the null distribution. We reject H_0 and conclude there is a linear relationship between weeks of gestation and birth weight.

20.5.3 Mathematical model: the t-test for slope

When the technical conditions are met, we can use the t-distribution to test the slope:

$$T = \frac{b_1 - 0}{SE_{b_1}}$$

This T-statistic follows a t-distribution with $df = n - 2$.

The regression output for predicting the change in House seats for the President's party from the unemployment rate is:

Term	Estimate	Std. Error	T statistic	p-value
(Intercept)	-7.36	5.16	-1.43	0.1649
unemp	-0.89	0.835	-1.07	0.2961

The p-value of 0.2961 is not discernible. The data do not provide convincing evidence that the unemployment rate is a useful predictor of midterm election outcomes.

20.5.4 Confidence intervals for the slope

Confidence intervals for model coefficients.

Confidence intervals for model coefficients can be computed using the t-distribution:

$$b_1 \pm t^*_{df} \times SE_{b_1}$$

where t^*_{df} is the critical value corresponding to the confidence level, with $df = n - 2$.

Using the Elmhurst College regression output ($b_1 = -0.0431$, $SE = 0.0108$, $df = 48$), compute a 95% confidence interval for the slope.

The critical value is $t^*_{48} = 2.01$. The confidence interval is:

$$-0.0431 \pm 2.01 \times 0.0108 = (-0.0648, -0.0214)$$

We are 95% confident that for each additional \$1,000 in family income, the university's gift aid decreases by between \$21.40 and \$64.80 on average.

20.6 Conditions for linear regression

For the mathematical inference to be valid, we check four conditions, often remembered by the **LINE** mnemonic:

- **L – Linear model:** The data should show a linear trend. Check the scatterplot and residual plot.
- **I – Independent observations:** Be cautious with time series data or clustered observations.
- **N – Nearly normal residuals:** Residuals should be approximately normally distributed. This is less critical for large samples (Central Limit Theorem), but outliers are always a concern.
- **E – Equal variability:** The variability of points around the line should be roughly constant across all values of x. A "fan shape" in the residual plot indicates a violation.

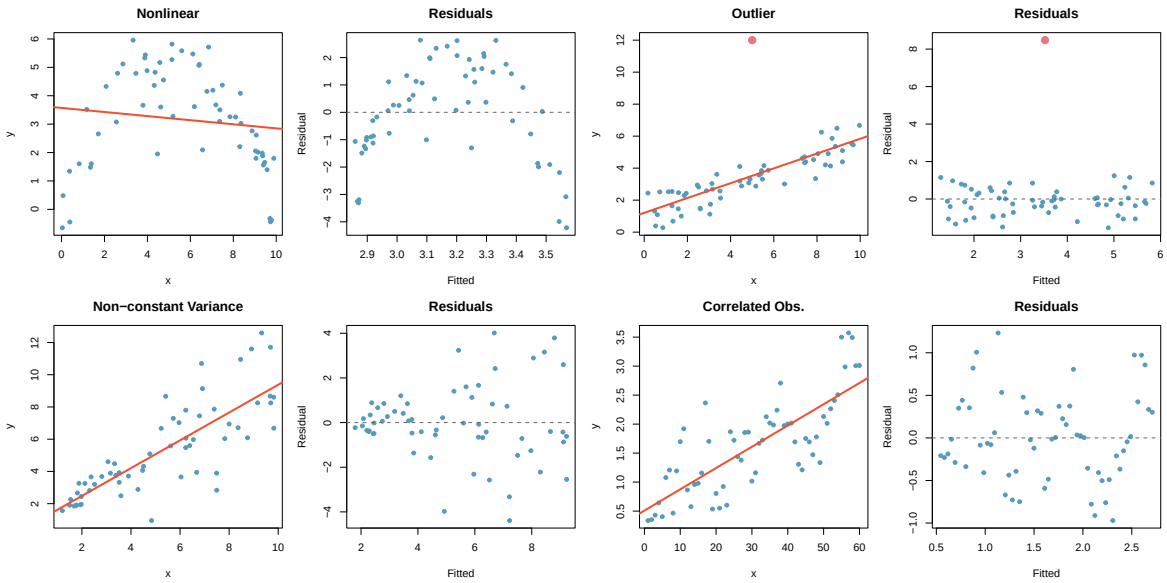


Figure 20.6: Four scatterplots and their residual plots showing common violations: nonlinearity, an outlier, increasing variability, and correlated observations.

Diagnostics for linear regression.

Independence is always important. The normality condition matters most for small samples. The constant variance condition is especially important for inference. The linearity condition is the most fundamental – if the true relationship is not linear, the slope estimate has no meaningful interpretation.

20.7 Chapter review

20.7.1 Summary

In this chapter, we developed the least squares regression line as a tool for modeling the linear relationship between two quantitative variables. The slope describes the predicted change in y for a one-unit increase in x, and the intercept describes the predicted value of y when $x = 0$. Residuals measure the difference between observed and predicted values, and residual plots help assess whether the linear model is appropriate. The coefficient of determination R^2 quantifies the proportion of variability in y explained by the model. For inference, we use either a randomization test or the t-distribution to test whether the slope is discernibly different from zero, and we can construct confidence intervals for the slope. The LINE conditions (linearity, independence, normality, equal variance) must be checked for the mathematical model to be valid.

20.7.2 Key terms

- Predictor / outcome variable
- Least squares line
- Slope (b_1) and intercept (b_0)
- Residual
- Residual plot
- R-squared (R^2) / coefficient of determination
- Total sum of squares (SST) / sum of squared errors (SSE)

- T-test for slope
- LINE conditions
- Extrapolation

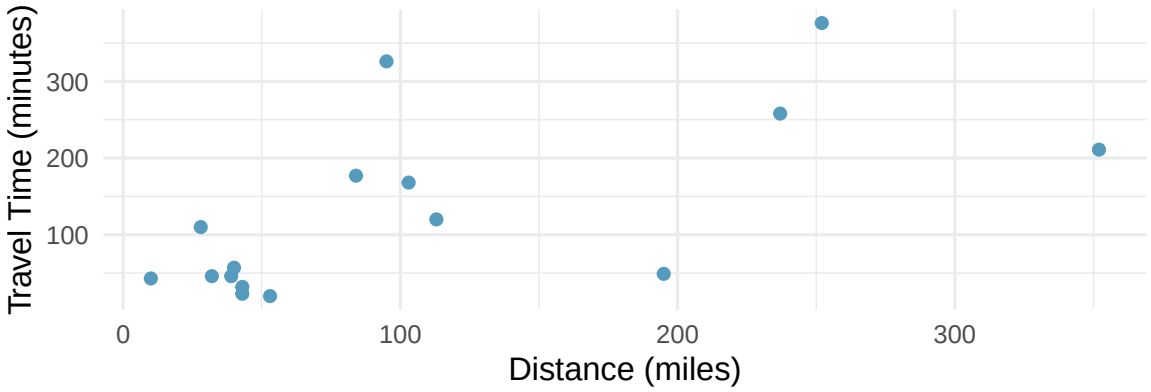
20.8 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Units of regression.** Consider a regression predicting the number of calories (cal) from width (cm) for a sample of square shaped chocolate brownies. What are the units of the correlation coefficient, the intercept, and the slope?

2. **Which is higher?** Determine if (I) or (II) is higher or if they are equal: "*For a regression line, the uncertainty associated with the slope estimate, b_1, is higher when (I) there is a lot of scatter around the regression line or (II) there is very little scatter around the regression line.*" Explain your reasoning.

3. **The Coast Starlight, regression.** The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes). The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.



   a. Write the equation of the regression line for predicting travel time.

   b. Interpret the slope and the intercept in this context.

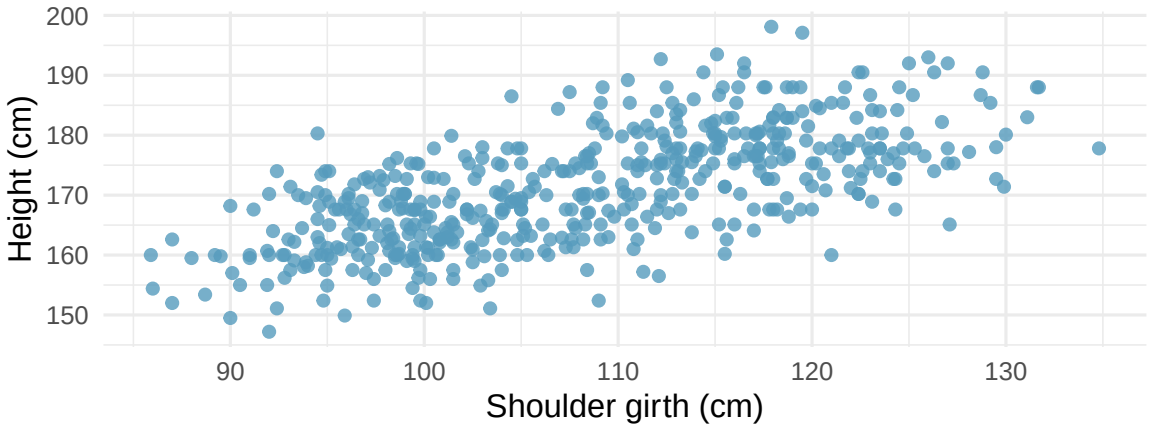
   c. Calculate R^2 of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret R^2 in the context of the application.

   d. The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.

   e. It takes the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.

   f. Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

4. **Body measurements, regression.** Researchers studying anthropometry collected body and skeletal diameter measurements, as well as age, weight, height and sex for 507 physically active individuals. The scatterplot below shows the relationship between height and shoulder girth (circumference of shoulders measured over deltoid muscles), both measured in centimeters. The mean shoulder girth is 107.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shoulder girth is 0.67. (Heinz et al. 2003)



   a. Write the equation of the regression line for predicting height.

   b. Interpret the slope and the intercept in this context.

   c. Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.

   d. A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.

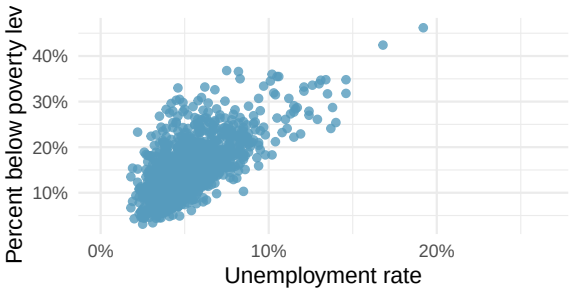
   e. The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means.

   f. A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

5. **Poverty and unemployment.** The following scatterplot shows the relationship between percent of population below the poverty level (`poverty`) from unemployment rate among those ages 20-64 (`unemployment_rate`) in counties in the US, as provided by data from the 2019 American Community Survey. The regression output for the model for predicting `poverty` from `unemployment_rate` is also provided.

term	estimate	std.error	statistic	p.value
`(Intercept)`	4.60	0.349	13.2	<0.0001
`unemployment_rate`	2.05	0.062	33.1	<0.0001

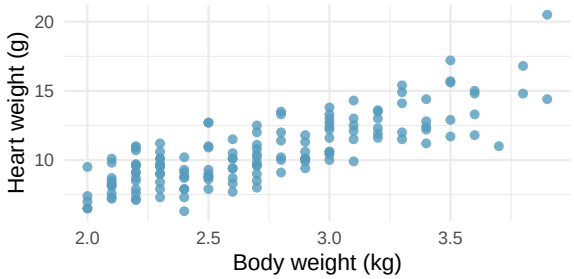
a. Write out the linear model.
b. Interpret the intercept.
c. Interpret the slope.
d. The R^2 of this model is 46%. Interpret this value.
e. Calculate the correlation coefficient.



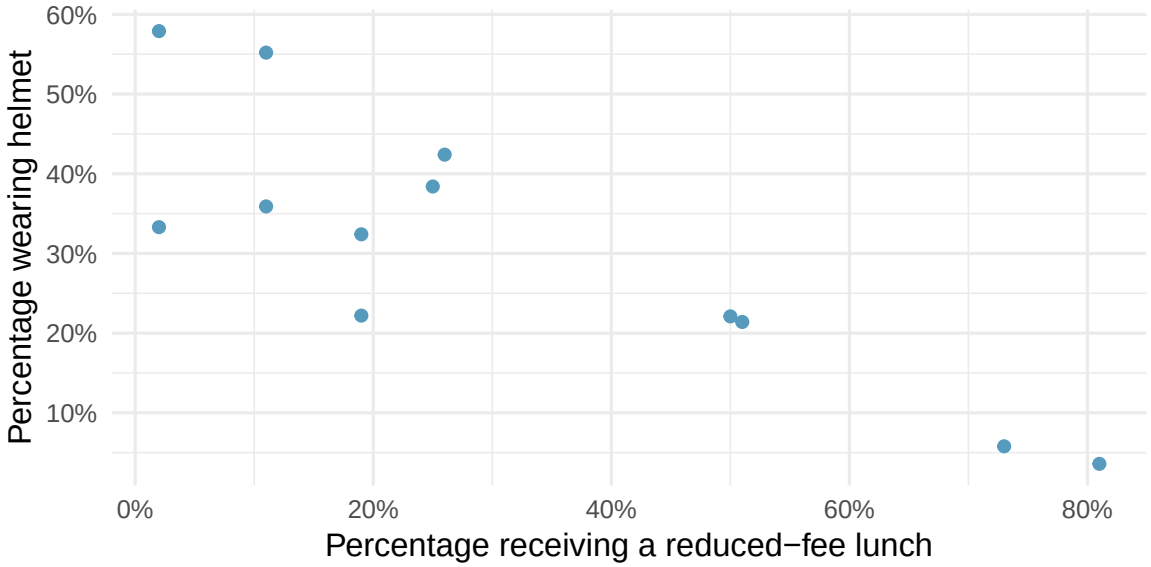
6. **Cat weights.** The following regression output is for predicting the heart weight (`Hwt`, in g) of cats from their body weight (`Bwt`, in kg). The coefficients are estimated using a dataset of 144 domestic cats.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-0.357	0.692	-0.515	0.6072
`Bwt`	4.034	0.250	16.119	<0.0001

a. Write out the linear model.
b. Interpret the intercept.
c. Interpret the slope.
d. The R^2 of this model is 65%. Interpret R^2.
e. Calculate the correlation coefficient.



7. **Helmets and lunches.** The scatterplot shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (`lunch`) and the percentage of bike riders in the neighborhood wearing helmets (`helmet`). The average percentage of children receiving reduced-fee lunches is 30.83% with a standard deviation of 26.72% and the average percentage of bike riders wearing helmets is 30.88% with a standard deviation of 16.95%.

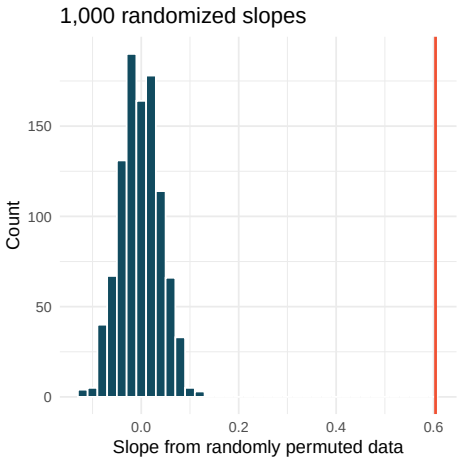


a. If the R^2 for the least-squares regression line for these data is 72%, what is the correlation between `lunch` and `helmet`?
b. Calculate the slope and intercept for the least-squares regression line for these data.
c. Interpret the intercept of the least-squares regression line in the context of the application.
d. Interpret the slope of the least-squares regression line in the context of the application.

e. What would the value of the residual be for a neighborhood where 40% of the children receive reduced-fee lunches and 40% of the bike riders wear helmets? Interpret the meaning of this residual in the context of the application.

8. **Body measurements, randomization test.** Researchers studying anthropometry collected body and skeletal diameter measurements, as well as age, weight, height and sex for 507 physically active individuals. A linear model is built to predict height based on shoulder girth (circumference of shoulders measured over deltoid muscles), both measured in centimeters. (Heinz et al. 2003) Shown below are the linear model output for predicting height from shoulder girth and the histogram of slopes from 1,000 randomized datasets (1,000 times, `hgt` was permuted and regressed against `sho_gi`). The red vertical line is drawn at the observed slope value which was produced in the linear model output.

term	estimate	std.error	statistic	p.value
`(Intercept)`	105.832	3.27	32.3	<0.0001
`sho_gi`	0.604	0.03	20.0	<0.0001



a. What are the null and alternative hypotheses for evaluating whether the slope of the model predicting height from shoulder girth is differen than 0.

b. Using the histogram which describes the distribution of slopes when the null hypothesis is true, find the p-value and conclude the hypothesis test in the context of the problem (use words like shoulder girth and height).

c. Is the conclusion based on the histogram of randomized slopes consistent with the conclusion from the mathematical model? Explain your reasoning.

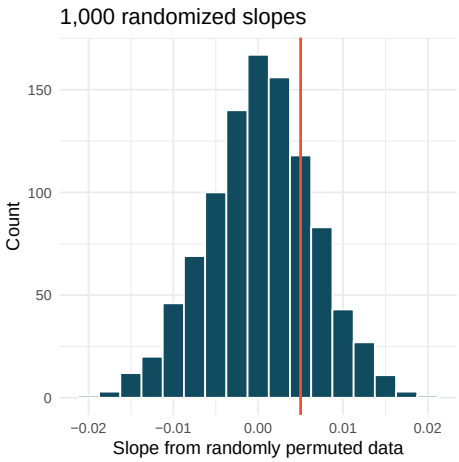
9. **Baby's weight and father's age, randomization test.** US Department of Health and Human Services, Centers for Disease Control and Prevention collect information on births recorded in the country. The data used here are a random sample of 1000 births from 2014. Here, we study the relationship between the father's age and the weight of the baby. (ICPSR 2014) Shown below are the linear model output for predicting baby's weight (in pounds) from father's age (in years) and the histogram of slopes from 1000 randomized datasets (1000 times, `weight` was permuted and regressed against `fage`). The red vertical line is drawn at the observed slope value which was produced in the linear model output.

a. What are the null and alternative hypotheses for evaluating whether the slope of the model for predicting baby's weight from father's age is different than 0?

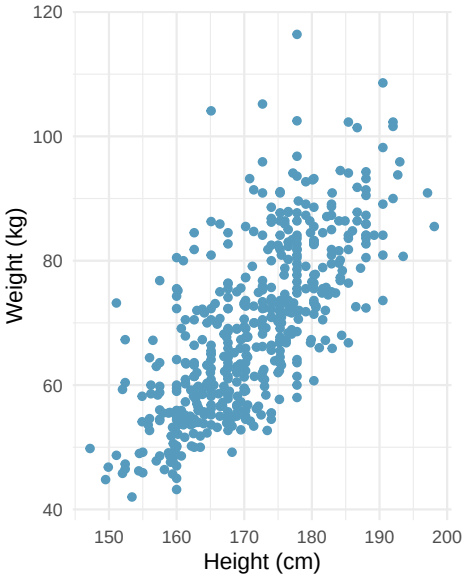
b. Using the histogram which describes the distribution of slopes when the null hypothesis is true, find the p-value and conclude the hypothesis test in the context of the problem (use words like father's age and weight of baby). What does the conclusion of your test say about whether the father's age is a useful predictor of baby's weight?

c. Is the conclusion based on the histogram of randomized slopes consistent with the conclusion from the mathematical model? Explain your reasoning.

term	estimate	std.error	statistic	p.value
(Intercept)	7.101	0.199	35.674	<0.0001
fage	0.005	0.006	0.757	0.4495



10. **Body measurements, mathematical test.** The scatterplot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals. (Heinz et al. 2003)



term	estimate	std.error	statistic	p.value
(Intercept)	-105.01	7.54	-13.9	<0.0001
hgt	1.02	0.04	23.1	<0.0001

a. Describe the relationship between height and weight.

b. Write the equation of the regression line. Interpret the slope and intercept in context.

c. Do the data provide convincing evidence that the true slope parameter is different than 0? State the null and alternative hypotheses, report the p-value (using a mathematical model), and state your conclusion.

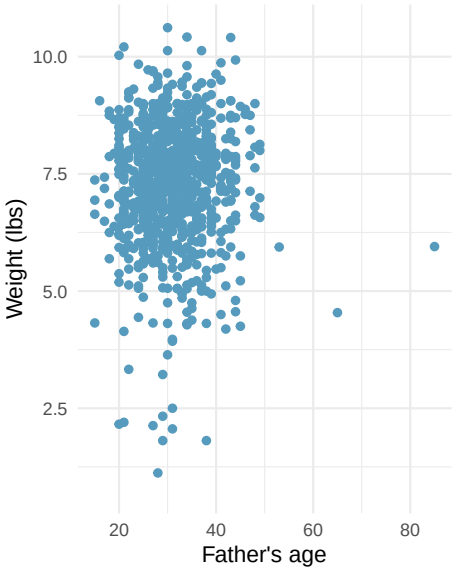
d. The correlation coefficient for height and weight is 0.72. Calculate R^2 and interpret it in context.

11. **Baby's weight and father's age, mathematical test.** Is the father's age useful in predicting the baby's weight? The scatterplot and least squares summary below show the relationship between baby's weight (measured in pounds) and father's age for a random sample of babies. (ICPSR 2014)

a. What is the predicted weight of a baby whose father is 30 years old?

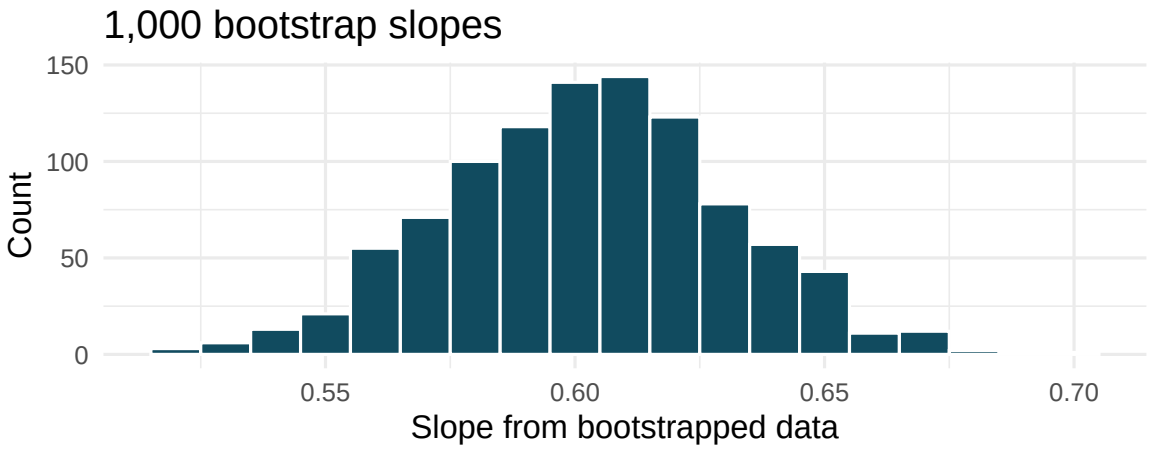
b. Do the data provide convincing evidence that the model for predicting baby weights from father's age has a slope different than 0? State the null and alternative hypotheses, report the p-value (using a mathematical model), and state your conclusion.

c. Based on your conclusion, is father's age a useful predictor of baby's weight?



term	estimate	std.error	statistic	p.value
`(Intercept)`	7.1042	0.1936	36.698	<0.0001
`fage`	0.0047	0.0061	0.779	0.4359

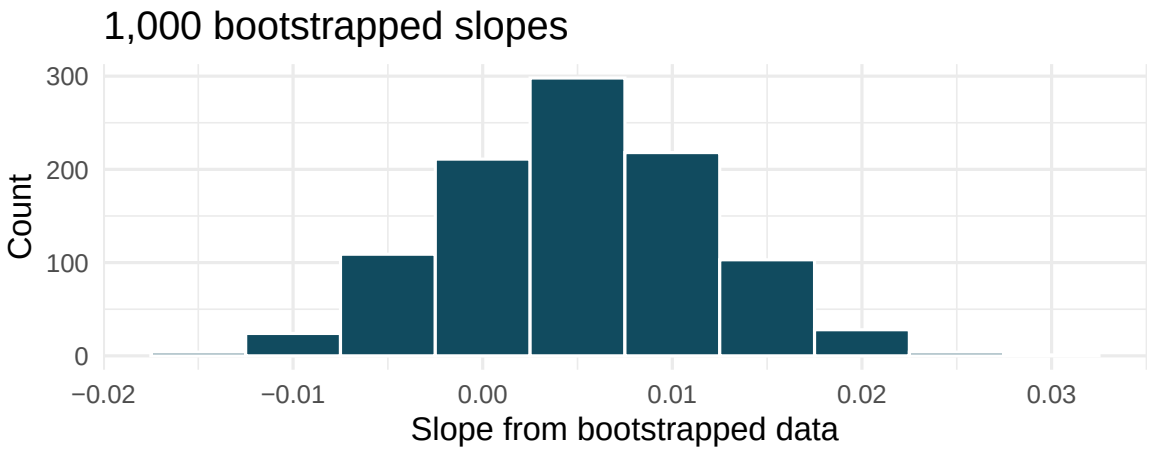
12. **Body measurements, bootstrap percentile interval.** In order to estimate the slope of the model predicting height based on shoulder girth (circumference of shoulders measured over deltoid muscles), 1,000 bootstrap samples are taken from a dataset of body measurements from 507 people. A linear model predicting height based on shoulder girth is fit to each bootstrap sample, and the slope is estimated. A histogram of these slopes is shown below. (Heinz et al. 2003)



a. Using the bootstrap percentile method and the histogram above, find a 98% confidence interval for the slope parameter.

b. Interpret the confidence interval in the context of the problem.

13. **Baby's weight and father's age, bootstrap percentile interval.** US Department of Health and Human Services, Centers for Disease Control and Prevention collect information on births recorded in the country. The data used here are a random sample of 1000 births from 2014. Here, we study the relationship between the father's age and the weight of the baby. Below is the bootstrap distribution of the slope statistic from 1,000 different bootstrap samples of the data. (ICPSR 2014)

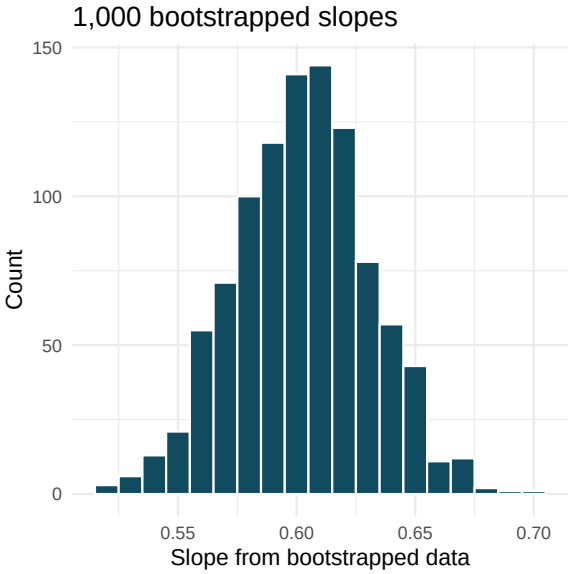


a. Using the bootstrap percentile method and the histogram above, find a 95% confidence interval for the slope parameter.

b. Interpret the confidence interval in the context of the problem.

14. **Body measurements, standard error bootstrap interval.** A linear model is built to predict height based on shoulder girth (circumference of shoulders measured over deltoid muscles), both measured in centimeters. (Heinz et al. 2003) Shown below are the linear model output for predicting height from shoulder girth and the bootstrap distribution of the slope statistic from 1,000 different bootstrap samples of the data.

term	estimate	std.error	statistic	p.value
`(Intercept)`	105.832	3.27	32.3	<0.0001
`sho_gi`	0.604	0.03	20.0	<0.0001

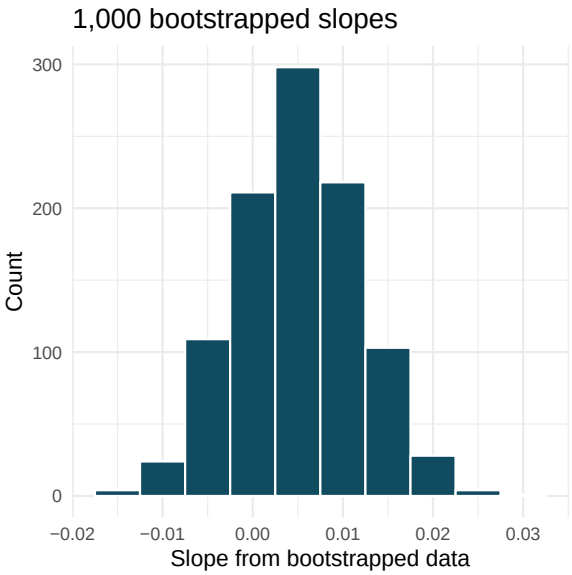


a. Using the histogram, approximate the standard error of the slope statistic (that is, quantify the variability of the slope statistic from sample to sample).
b. Find a 98% bootstrap SE confidence interval for the slope parameter.
c. Interpret the confidence interval in the context of the problem.

15. **Baby's weight and father's age, standard error bootstrap interval.** US Department of Health and Human Services, Centers for Disease Control and Prevention collect information on births recorded in the country. The data used here are a random sample of 1000 births from 2014. Here, we study the relationship between the father's age and the weight of the baby. (ICPSR 2014) Shown below are the linear model output for predicting baby's weight (in pounds) from father's age (in years) and the the bootstrap distribution of the slope statistic from 1000 different bootstrap samples of the data.

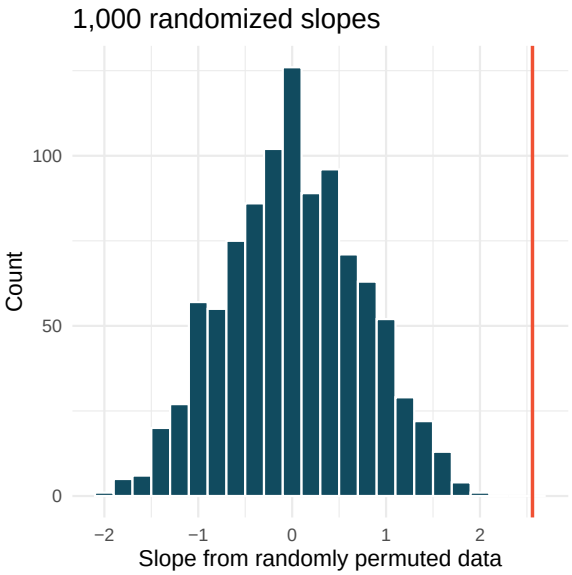
term	estimate	std.error	statistic	p.value
`(Intercept)`	7.101	0.199	35.674	<0.0001
`fage`	0.005	0.006	0.757	0.4495

a. Using the histogram, approximate the standard error of the slope statistic (that is, quantify the variability of the slope statistic from sample to sample).
b. Find a 95% bootstrap SE confidence interval for the slope parameter.
c. Interpret the confidence interval in the context of the problem.



16. **Murders and poverty, randomization test.** The following regression output is for predicting annual murders per million (`annual_murders_per_mil`) from percentage living in poverty (`perc_pov`) in a random sample of 20 metropolitan areas. Shown below are the linear model output for predicting annual murders per million from percentage living in poverty for metropolitan areas and the histogram of slopes from 1000 randomized datasets (1000 times, `annual_murders_per_mil` was permuted and regressed against `perc_pov`). The red vertical line is drawn at the observed slope value which was produced in the linear model output.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-29.90	7.79	-3.84	0.0012
`perc_pov`	2.56	0.39	6.56	<0.0001



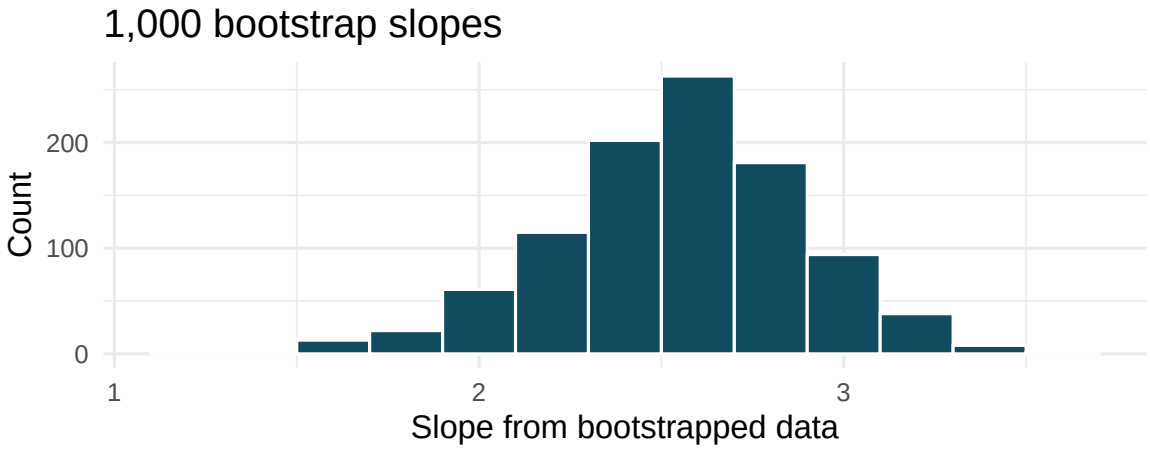
a. What are the null and alternative hypotheses for evaluating whether the slope of the model for predicting annual murder rate from poverty percentage is different than 0?
b. Using the histogram which describes the distribution of slopes when the null hypothesis is true, find the p-value and conclude the hypothesis test in the context of the problem (use words like murder rate and poverty).
c. Is the conclusion based on the histogram of randomized slopes consistent with the conclusion which would have been obtained using the mathematical model? Explain your reasoning.

17. **Murders and poverty, mathematical test.** The table below shows the output of a linear model annual murders per million (`annual_murders_per_mil`) from percentage living in poverty (`perc_pov`) in a random sample of 20 metropolitan areas.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-29.90	7.79	-3.84	0.0012
`perc_pov`	2.56	0.39	6.56	<0.0001

a. What are the hypotheses for evaluating whether the slope of the model predicting annual murder rate from poverty percentage is different than 0?
b. State the conclusion of the hypothesis test from part (a) in context. What does this say about whether poverty percentage is a useful predictor of annual murder rate?
c. Calculate a 95% confidence interval for the slope of poverty percentage, and interpret it in context.
d. Do your results from the hypothesis test and the confidence interval agree? Explain your reasoning.

18. **Murders and poverty, bootstrap percentile interval.** Data on annual murders per million (`annual_murders_per_mil`) and percentage living in poverty (`perc_pov`) is collected from a random sample of 20 metropolitan areas. Using these data we want to estimate the slope of the model predicting `annual_murders_per_mil` from `perc_pov`. We take 1,000 bootstrap samples of the data and fit a linear model predicting `annual_murders_per_mil` from `perc_pov` to each bootstrap sample. A histogram of these slopes is shown below.

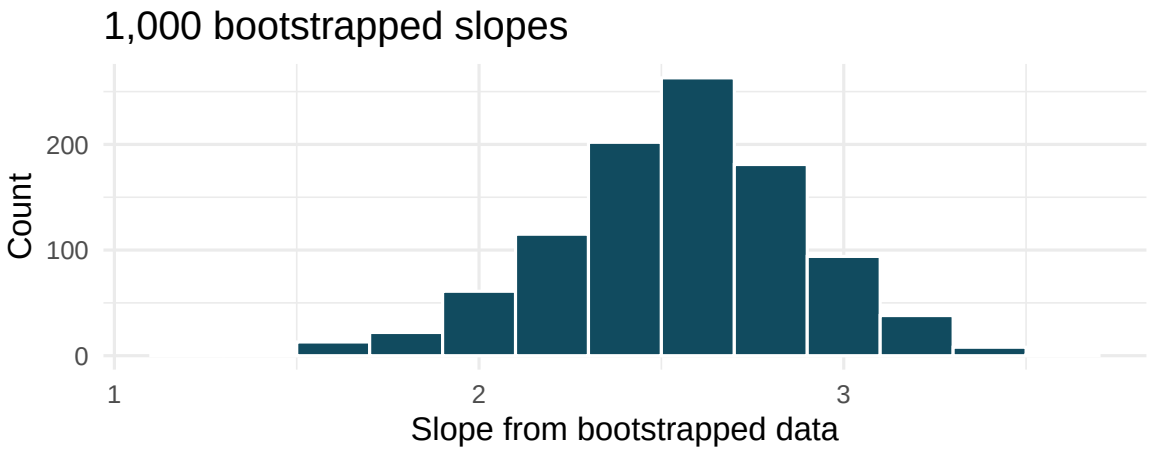


a. Using the percentile bootstrap method and the histogram above, find a 90% confidence interval for the slope parameter.

b. Interpret the confidence interval in the context of the problem.

19. **Murders and poverty, standard error bootstrap interval.** A linear model is built to predict annual murders per million (`annual_murders_per_mil`) from percentage living in poverty (`perc_pov`) in a random sample of 20 metropolitan areas. Shown below are the standard linear model output for predicting annual murders per million from percentage living in poverty for metropolitan areas and the bootstrap distribution of the slope statistic from 1000 different bootstrap samples of the data.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-29.90	7.79	-3.84	0.0012
`perc_pov`	2.56	0.39	6.56	<0.0001



a. Using the histogram, approximate the standard error of the slope statistic (that is, quantify the variability of the slope statistic from sample to sample).

b. Find a 90% bootstrap SE confidence interval for the slope parameter.

c. Interpret the confidence interval in the context of the problem.

20. **I heart cats.** Researchers collected data on heart and body weights of 144 domestic adult cats. The table below shows the output of a linear model predicting heart weight (measured in grams) from body weight (measured in kilograms) of these cats.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-0.357	0.692	-0.515	0.6072
`Bwt`	4.034	0.250	16.119	<0.0001

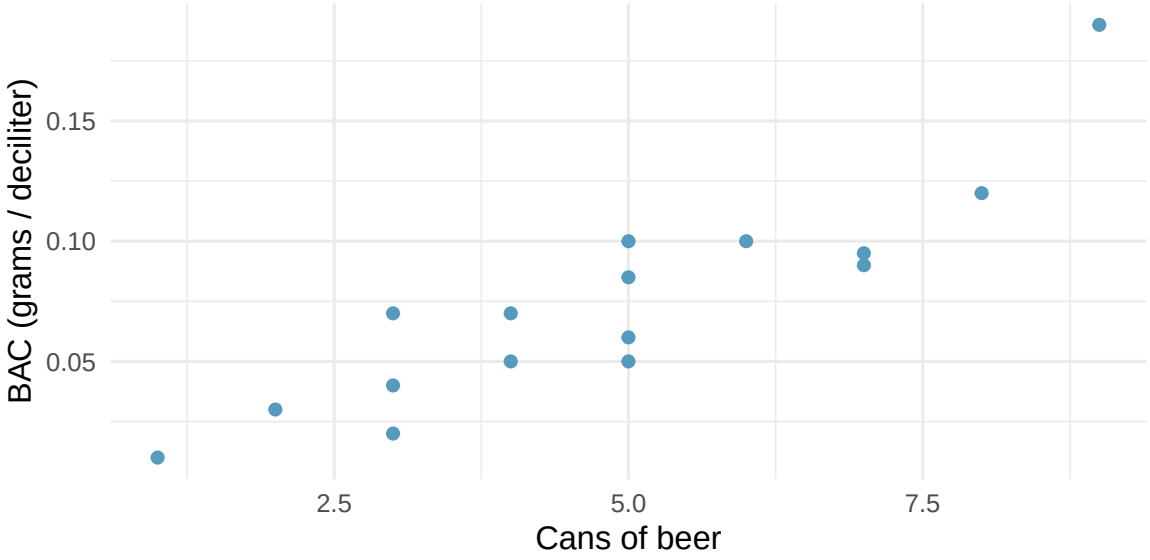
a. What are the hypotheses for evaluating whether body weight is positively associated with heart weight in cats?

b. State the conclusion of the hypothesis test from part (a) in context.

c. Calculate a 95% confidence interval for the slope of body weight, and interpret it in context.

d. Do your results from the hypothesis test and the confidence interval agree? Explain your reasoning.

21. **Beer and blood alcohol content.** Many people believe that weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person consumed. Here we examine data from sixteen student volunteers at Ohio State University who each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood. The scatterplot and regression table summarize the findings. (Malkevitch and Lesser 2008)



term	estimate	std.error	statistic	p.value
`(Intercept)`	-0.0127	0.0126	-1.00	0.332
`beers`	0.0180	0.0024	7.48	<0.0001

a. Describe the relationship between the number of cans of beer and BAC.

b. Write the equation of the regression line. Interpret the slope and intercept in context.

c. Do the data provide convincing evidence that drinking more cans of beer is associated with an increase in blood alcohol? State the null and alternative hypotheses, report the p-value, and state your conclusion.

d. The correlation coefficient for number of cans of beer and BAC is 0.89. Calculate R^2 and interpret it in context.

e. Suppose we visit a bar in our own town, ask people how many drinks they have had, and also measure their BAC. Would the relationship between number of drinks and BAC would be as strong as the relationship found in the Ohio State study? Why?

Datasets: `coast_starlight` (openintro) | `bdims` (openintro) | `county_2019` (usdata) | `cats` (MASS) | `births14` (openintro) | `cats` (MASS) | `bac` (openintro)

StatLens Exercises

These exercises focus on the parts of regression inference a calculator *can't* do for you: **framing the question**, **choosing the right procedure**, **checking the four regression conditions**, and **interpreting** slopes and intervals without falling into the standard traps. Let StatLens do the arithmetic.

1. **Frame the question.** A real-estate analyst fits a simple linear regression of house **price** (response, \$1000s) on **square footage** (explanatory) using 200 homes in one zip code. The fitted line is $\widehat{\text{price}} = 80 + 0.15 \cdot \text{sqft}$.
   a. Identify the explanatory and response variables and give the symbol for the **population slope**.
   b. Write the null and alternative hypotheses for a two-sided test of whether square footage is associated with price.
   c. The slope is 0.15. **Interpret it in context, including units.** Then identify the most common error students make in slope interpretation.

2. **Choose the procedure.** For each scenario, name the procedure — **CI for slope** β_1, **hypothesis test for** β_1, **correlation test** ($\rho = 0$), or **not a regression problem**.
   a. Is there evidence of *any* linear association between hours studied and exam score?
   b. Estimate, with a margin of error, how much exam score changes per extra hour studied.
   c. Compare the *mean* exam score in morning vs. evening sections of the same course.
   d. Decide whether the **strength** of the linear relationship between two variables is statistically discernible from zero.
   e. Estimate the mean weight of all puppies in a litter from a sample of 8.

3. **Check the conditions before you trust the slope.** Open a regression dataset in the Regression Slope tool; pair it with the Regression scatterplot for diagnostics.
   a. Name the **four** regression conditions (often summarized as "LINE": **L**inearity, **I**ndependence, **N**ormality of residuals, **E**qual variance of residuals).
   b. Open the residual plot. What pattern in the residuals would convince you that the **linearity** condition has failed? What pattern would convince you the **equal-variance** condition has failed?
   c. Suppose one observation has a residual five times larger than any other. Should you discard it before re-running the regression? What two questions should you ask first?

4. **Run it and interpret it.** Open any regression dataset in the Regression Slope tool.
   a. Report the estimated slope b_1, its standard error, the t statistic, and the p-value.
   b. Read off the 95% confidence interval for β_1. Does it contain 0?
   c. Write a one-sentence conclusion **in context**, naming both variables, the direction, and the strength of evidence.
   d. Which statement is the correct interpretation of the 95% CI for β_1?
      i. There is a 95% probability that the true slope lies in this interval.
      ii. If the study were repeated many times, about 95% of the intervals built this way would contain the true slope.
      iii. 95% of the data points fall within this interval.
      iv. About 95% of the predicted responses fall within this interval.

5. **Simulation vs. analytic — the bootstrap slope as a check.** The t-CI for the slope $(b_1 \pm t^* \cdot \text{SE}(b_1))$ assumes the LINE conditions hold. The bootstrap CI for the slope assumes only that the data are an i.i.d. sample from the population. You will compare them.

   First a well-behaved case. Open any roughly linear, homoscedastic regression dataset in both the Regression Slope tool *and* the Bootstrap Slope tool (generate 5000 resamples in the latter).

   a. Compare the two 95% intervals for β_1. How close are they?

   Now a troublesome case. Pick a regression dataset (or paste in one) with **visible heteroscedasticity** — a clear fan-shaped residual pattern.

   b. Compare the two 95% intervals again. Do they still agree? Which one is **wider**?
   c. Explain the pattern. The bootstrap resamples whole (x, y) pairs, so it "sees" the heteroscedasticity — the resampled slopes fluctuate more in some parts of x than others. The t-formula uses a *single* SE pooled across all x, which under-counts the variability in the high-spread region. Which interval is more honest?
   d. A student concludes: "If the conditions fail, you must transform the data." Why is that the *first* impulse, but not the only one? When is **the bootstrap CI the better tool** even after a transformation?

Chapter 21

Prediction and Model Fit

In the previous chapter, we developed the least squares regression line, interpreted its slope and intercept, and performed inference to test whether a linear relationship exists. In this chapter, we focus on using the regression model for prediction and assessing how well it fits the data. We discuss the dangers of extrapolation, introduce residual diagnostics for identifying problems with the model, examine leverage and influential points, and briefly introduce confidence intervals and prediction intervals for forecasts.

21.1 Making predictions

One of the primary uses of a regression model is to predict the value of the response variable y for a given value of the explanatory variable x. To make a prediction, we simply plug the value of x into the regression equation.

The least squares line for predicting gift aid from family income at Elmhurst College is:

$$\widehat{\text{aid}} = 24.3 - 0.0431 \times \text{family\_income}$$

Predict the gift aid for a student whose family income is \$80,000 (i.e., $x = 80$ in \$1,000s).

$$\widehat{\text{aid}} = 24.3 - 0.0431 \times 80 = 24.3 - 3.45 = 20.85$$

The model predicts this student will receive approximately \$20,850 in gift aid.

Suppose a high school senior is considering Elmhurst College. Can they simply use the linear equation to calculate their financial aid?

They may use it as an estimate, but with important qualifiers. First, the data come from a single year, and aid policies may change. Second, the equation provides an imperfect estimate – while the linear model captures the overall trend, individual students' aid will vary around the predicted value.

21.2 Extrapolation

Use the Elmhurst model to estimate the aid for a student whose family has an income of \$1 million (i.e., $x = 1000$).

$$\widehat{\text{aid}} = 24.3 - 0.0431 \times 1000 = -18.8$$

The model predicts $-\$18,800$ in aid! Elmhurst College does not charge extra tuition to wealthy students. This nonsensical prediction illustrates the danger of **extrapolation**.

Extrapolation is treacherous.

Extrapolation means applying a model to make predictions outside the range of the observed data. A linear model is only an approximation of the true relationship, and that approximation may break down badly outside the range of the data used to build it.

Generally, predictions should only be made within the range of the observed predictor values. Predictions far outside this range are unreliable and should be treated with great skepticism.

Before making a prediction, always check whether the value of x falls within the range of the data used to build the model. If it does not, acknowledge that you are extrapolating and that the prediction may be unreliable.

21.3 Residual diagnostics

In the previous chapter, we introduced residual plots as a way to assess whether a linear model is appropriate. In this section, we examine residual diagnostics in more detail.

21.3.1 Patterns in residual plots

A well-fitting linear model produces residuals that appear randomly scattered around zero. Systematic patterns in the residuals indicate problems:

Consider the three pairs of scatterplots and residual plots below.

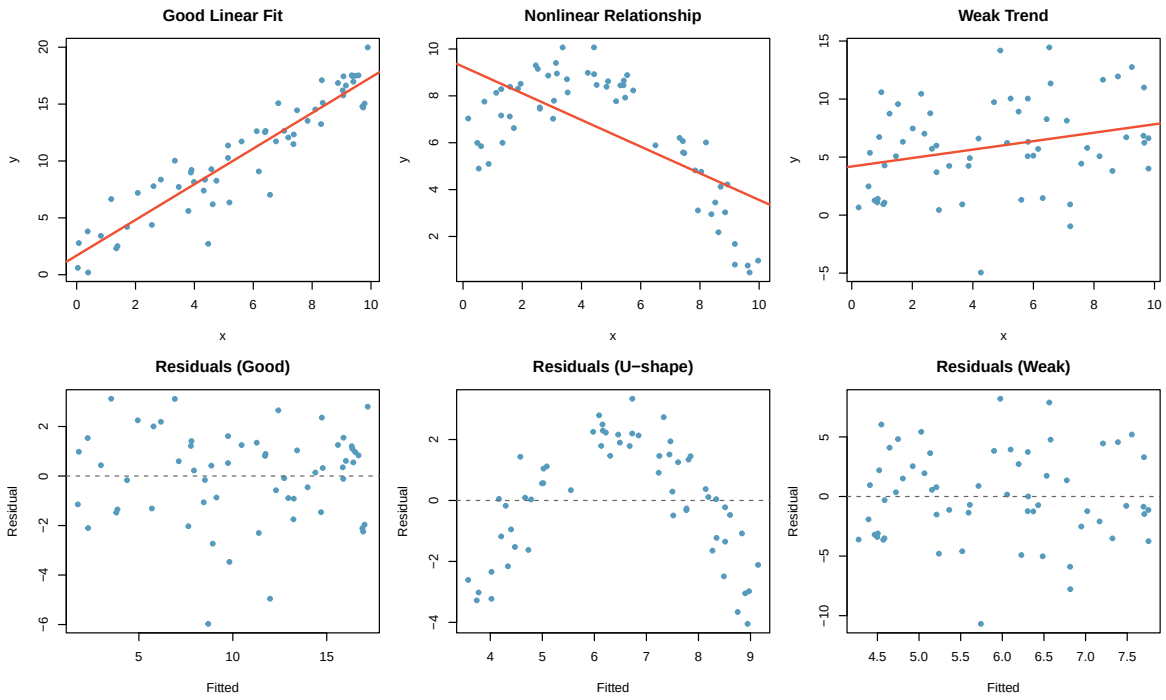


Figure 21.1: Three scatterplots with linear models in the top row and their corresponding residual plots in the bottom row. The first shows a good linear fit with random residuals. The second shows a curved relationship with a U-shaped residual pattern. The third shows a weak trend with random residuals. ↗ Try this live — toggle scatterplot ↔ residual plot and see the pattern each fit leaves behind.

Describe any patterns in the residuals.

Dataset 1: The residuals show no obvious patterns. They are scattered randomly around zero. The linear model is appropriate.

Dataset 2: There is a curved pattern in the residuals – they are negative at the edges and positive in the middle. This indicates a nonlinear relationship. A straight line is not a good model for these data.

Dataset 3: The residuals show no patterns, but the trend is very weak. The linear model is technically appropriate, but the relationship may not be practically meaningful.

21.3.2 Non-constant variability (heteroscedasticity)

When the spread of the residuals changes as the predicted value increases, we have **non-constant variability** (also called **heteroscedasticity**). This often appears as a "fan shape" in the residual plot – residuals spread out (or narrow) as the predicted values increase.

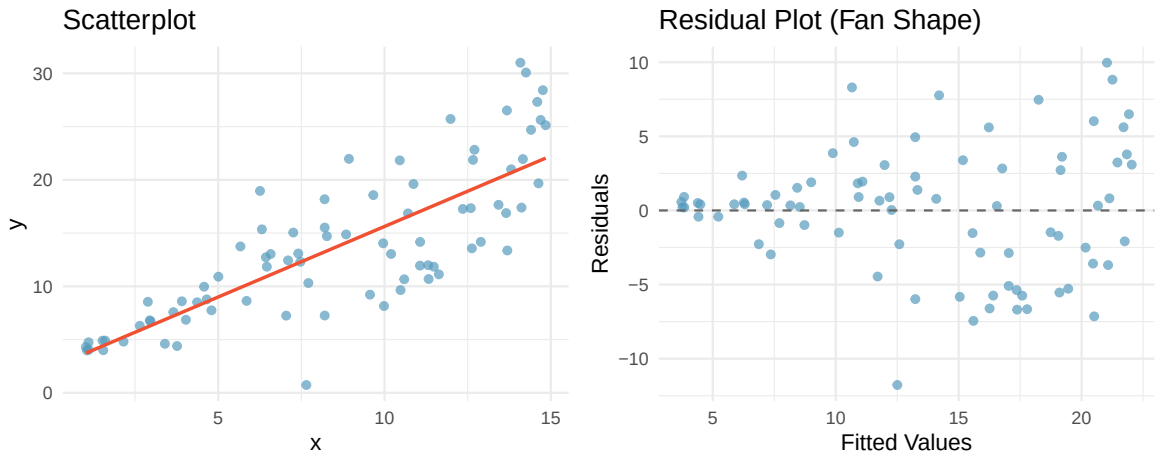


Figure 21.2: A scatterplot and residual plot showing increasing variability as x increases. The residual plot shows a fan shape. ↗ Try this live — the Ames home prices (sale price vs. living area) fan out, and the residual plot widens as x grows.

Non-constant variability does not bias the regression line itself, but it does affect inference: standard errors, confidence intervals, and p-values may be unreliable.

21.4 Leverage and influential points

Not all observations have the same impact on the regression line. Some points, by virtue of their position, can exert a disproportionate influence on the model.

21.4.1 Leverage

Leverage.

A point has **high leverage** if its x-value is far from $\bar{x}$ (the mean of the predictor variable). High-leverage points have the *potential* to strongly influence the slope and intercept of the regression line.

High leverage does not automatically mean the point is problematic. If a high-leverage point falls close to the trend established by the other data, it may actually strengthen the model. A high-leverage point is only problematic if its y-value does not follow the pattern of the rest of the data.

21.4.2 Influential points

Influential point.

A point is **influential** if removing it would substantially change the regression line (i.e., the estimated slope or intercept). Influential points are typically high-leverage points that do not follow the pattern of the rest of the data.

Consider three scenarios:

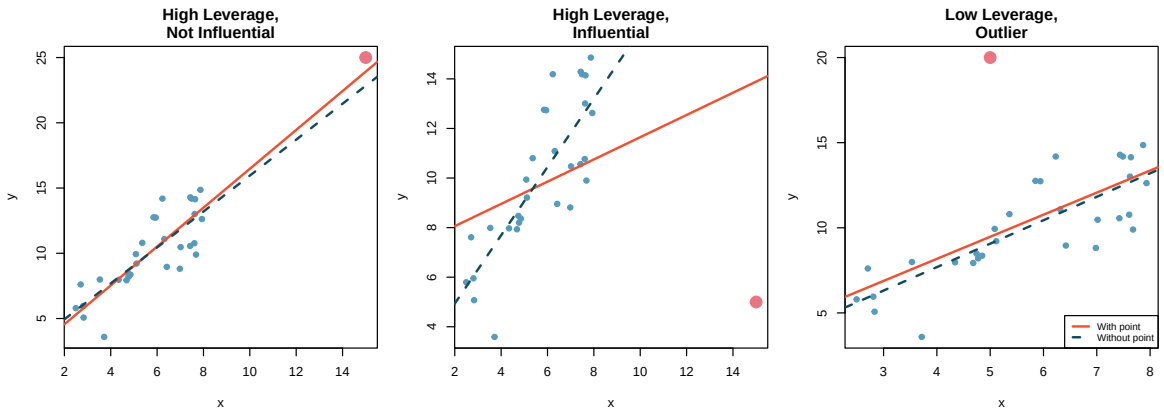


Figure 21.3: Three scatterplots with regression lines. In the first, a point in the corner has high leverage but follows the trend (not influential). In the second, a point far from the trend has high leverage and is influential – removing it changes the slope substantially. In the third, a point near the mean of x is an outlier in the y-direction but has low leverage and is not very influential.

Scenario 1: The point in the far right corner has high leverage (it is far from $\bar{x}$), but it falls along the trend of the other data. Removing it would barely change the regression line. It is *not* influential.

Scenario 2: The point in the far right corner has high leverage *and* falls far from the trend. Removing it would substantially change the slope. It *is* influential.

Scenario 3: The point near the center of the x-values is an outlier in the y-direction. It has low leverage because its x-value is near $\bar{x}$. Removing it would change the intercept slightly but not the slope very much. It has limited influence.

Dealing with influential points.

When you identify an influential point:

1. Verify that it is not a data entry error.
2. Run the analysis with and without the point and report both results.
3. Do not automatically remove it – if it is a legitimate observation, it contains real information.
4. Be transparent about its effect on the conclusions.

See it in action. Regression Explorer with CI + PI bands — toggles the 95% confidence band (mean response — the hourglass) and the wider 95% prediction band. For numeric bounds at a specific x_0, use the Slope tool point interval: type x_0 and get $\hat{y}$, the CI for the mean, and the prediction interval (with extrapolation warnings).

21.5 Confidence and prediction intervals

Extended Content. A standard one-semester introductory course stops the regression story before formal confidence and prediction intervals. This section is retained for institutions and instructors who want to cover them, and for students who want the deeper treatment. It reads as a self-contained section on top of the residual-diagnostics material above.

When we use a regression model to make predictions, there are two types of uncertainty to consider:

21.5.1 Confidence intervals for the mean response

A **confidence interval for the mean response** estimates the average value of y for all individuals with a particular value of x. This interval captures the uncertainty in estimating the population regression line.

For example, we might ask: "What is the average gift aid for all Elmhurst students with a family income of \$80,000?" The confidence interval for the mean response at $x = 80$ would give a range of plausible values for this population average.

The formula for a $(1 - \alpha) \times 100\%$ confidence interval for the mean response at $x = x^*$ is:

$$\hat{y} \pm t^*_{n-2} \cdot SE_{\hat{y}}$$

where the standard error depends on the distance of x^* from $\bar{x}$:

$$SE_{\hat{y}} = s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Here $s_e = \sqrt{MSE}$ is the residual standard error.

21.5.2 Prediction intervals for individual observations

A **prediction interval** estimates the value of y for a *single* new individual with a particular value of x. This interval is always wider than the confidence interval for the mean response because it must account for both:

1. The uncertainty in estimating the regression line (same as the confidence interval).
2. The natural variability of individual observations around the line.

The prediction interval at $x = x^*$ is:

$$\hat{y} \pm t^*_{n-2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Notice the extra "1+" under the square root compared to the confidence interval formula. This is what makes the prediction interval wider.

21.5.3 Comparing the two intervals

For the possum data, suppose we want to predict the head length for possums with a total length of 85 cm.

The least squares fit computed by software is $\hat{y} = 42.71 + 0.573x$, so the point prediction is $\hat{y} = 91.41$ mm. (The rounded line $\hat{y} = 41 + 0.59x$ from the previous chapter gives nearly the same prediction, 91.15 mm.)

- A 95% **confidence interval for the mean** head length of all possums with total length 85 cm is (90.84, 91.97) mm.
- A 95% **prediction interval** for the head length of a single possum with total length 85 cm is (86.23, 96.58) mm.

The confidence interval is relatively narrow because it estimates an average, which is estimated with moderate precision. The prediction interval is much wider because it must also account for the natural variation among individual possums – even possums with the same total length have different head lengths.

Confidence interval vs. prediction interval.

- A **confidence interval for the mean response** estimates the average y for a given x. It accounts for uncertainty in the regression line.
- A **prediction interval** estimates the y for a single new observation at a given x. It accounts for both the uncertainty in the regression line and the variability of individual observations.
- Prediction intervals are always wider than confidence intervals.
- Both intervals are narrowest at $x = \bar{x}$ and widen as x moves away from $\bar{x}$.

The fact that both intervals widen as x moves away from $\bar{x}$ connects back to the danger of extrapolation: predictions become increasingly uncertain farther from the center of the data.

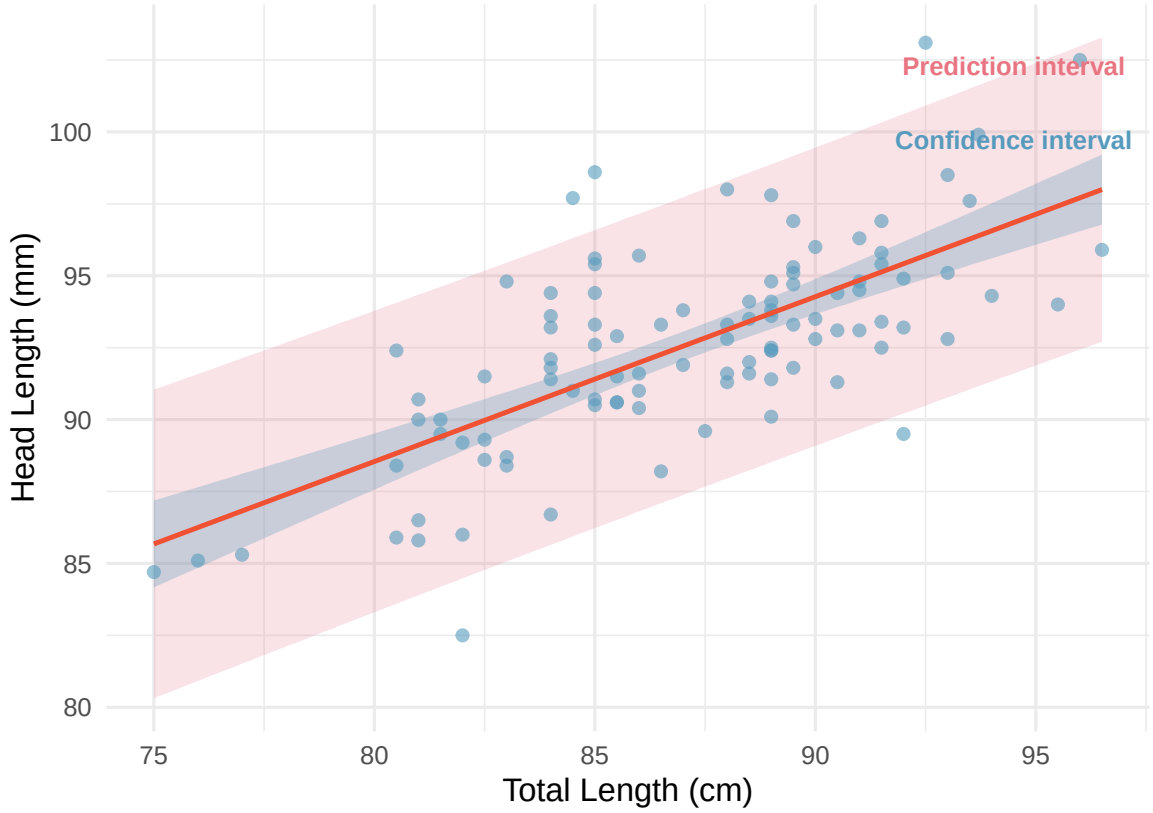


Figure 21.4: A scatterplot with a regression line and both 95% confidence bands (narrow) and 95% prediction bands (wide). Both bands are narrowest near the mean of x and flare outward. Open in StatLens

21.6 Chapter review

21.6.1 Summary

In this chapter, we focused on using regression models for prediction and evaluating their fit. We learned that extrapolation – predicting outside the range of the observed data – is dangerous and can

lead to nonsensical predictions. Residual diagnostics help us identify violations of the model assumptions, including nonlinearity, non-constant variance, and the influence of unusual points. Leverage measures how far a point's x-value is from the center of the data; influential points are high-leverage observations that do not follow the overall trend and substantially affect the regression line when removed. Finally, we distinguished between confidence intervals for the mean response and prediction intervals for individual observations, noting that prediction intervals are always wider because they account for individual variability.

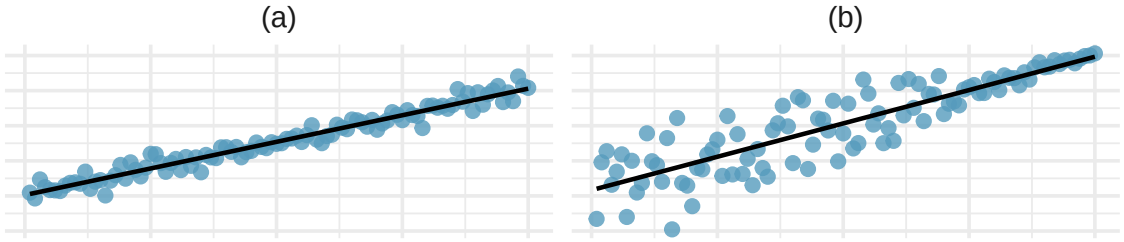
21.6.2 Key terms

- Extrapolation
- Residual diagnostics
- Non-constant variability (heteroscedasticity)
- Leverage
- Influential point
- Confidence interval for the mean response
- Prediction interval

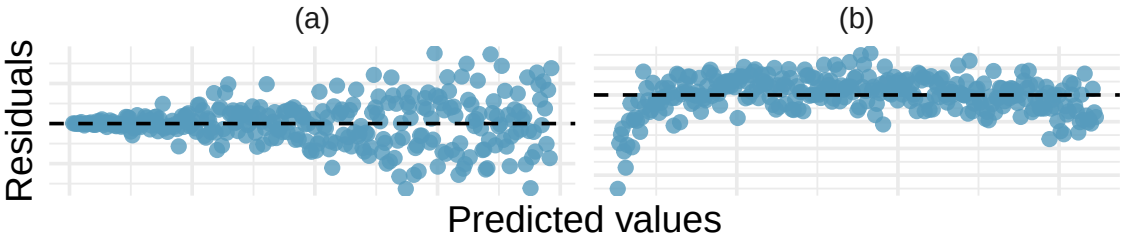
21.7 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Visualizing residuals.** The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe in words what those plots would look like.



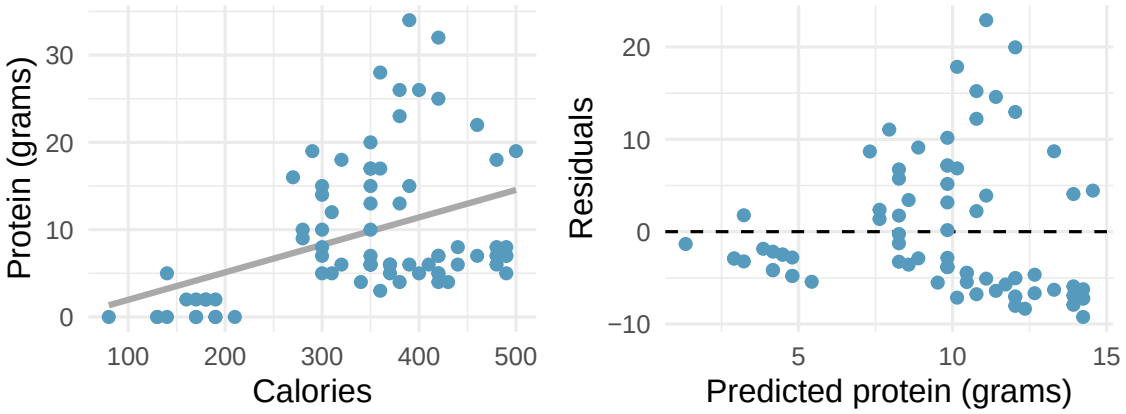
2. **Trends in residuals.** Shown below are two plots of residuals remaining after fitting a linear model to two different sets of data. For each plot, describe important features and determine if a linear model would be appropriate for these data. Explain your reasoning.



3. **Over-under, I.** Suppose we fit a regression line to predict the shelf life of an apple based on its weight. For a particular apple, we predict the shelf life to be 4.6 days. The apple's residual is -0.6 days. Did we over or under estimate the shelf-life of the apple? Explain your reasoning.

4. **Over-under, II.** Suppose we fit a regression line to predict the number of incidents of skin cancer per 1,000 people from the number of sunny days in a year. For a particular year, we predict the incidence of skin cancer to be 1.5 per 1,000 people, and the residual for this year is 0.5. Did we over or under estimate the incidence of skin cancer? Explain your reasoning.

5. **Starbucks, calories, and protein.** The scatterplot below shows the relationship between the number of calories and amount of protein (in grams) Starbucks food menu items contain. Since Starbucks only lists the number of calories on the display items, we might be interested in predicting the amount of protein a menu item has based on its calorie content.



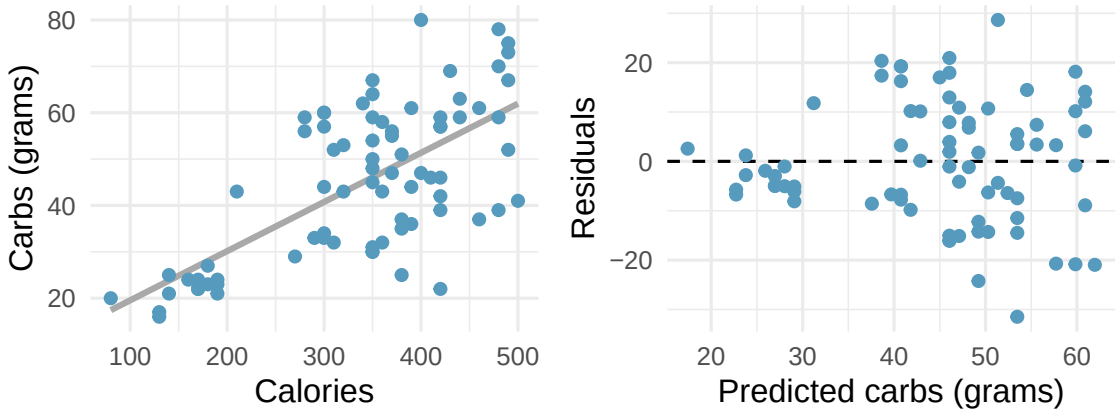
a. Describe the relationship between number of calories and amount of protein (in grams) that Starbucks food menu items contain.

b. In this scenario, what are the predictor and outcome variables?

c. Why might we want to fit a regression line to these data?

d. What does the residuals vs. predicted plot tell us about the variability in prediction errors based on this model for items with lower vs. higher predicted protein?

6. **Starbucks, calories, and carbs.** The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain. Since Starbucks only lists the number of calories on the display items, we might be interested in predicting the amount of carbs a menu item has based on its calorie content.



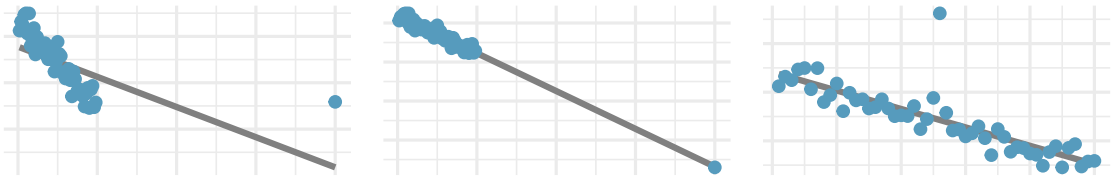
a. Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.

b. In this scenario, what are the predictor and outcome variables?

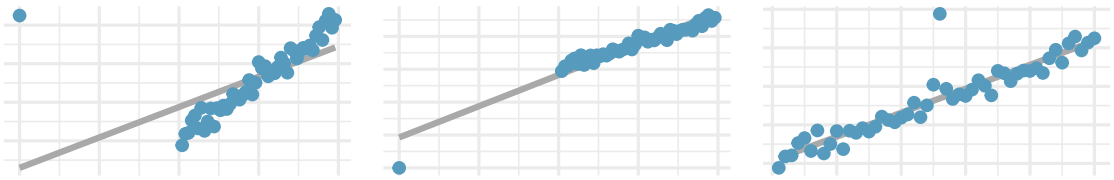
c. Why might we want to fit a regression line to these data?

d. What does the residuals vs. predicted plot tell us about the variability in prediction errors based on this model for items with lower vs. higher predicted carbs?

7. **Outliers, I.** Identify the outliers in the scatterplots shown below, and determine what type of outliers they are. Explain your reasoning.

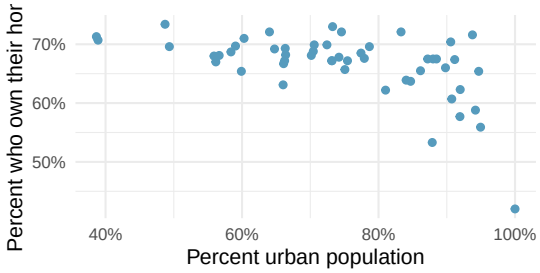


8. **Outliers, II.** Identify the outliers in the scatterplots shown below and determine what type of outliers they are. Explain your reasoning.

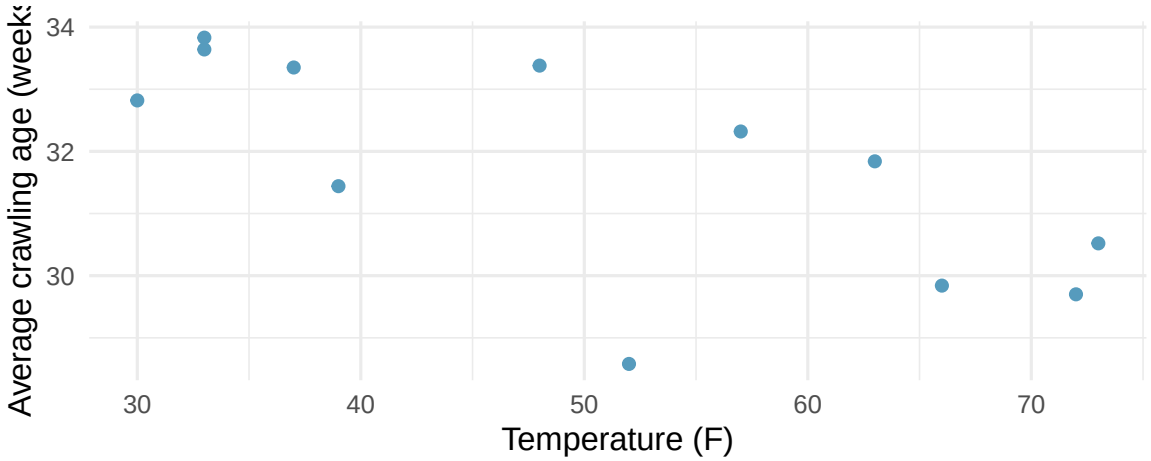


9. **Urban homeowners, outliers.** The scatterplot below shows the percent of families who own their home vs. the percent of the population living in urban areas. There are 52 observations, each corresponding to a state in the US. Puerto Rico and District of Columbia are also included.

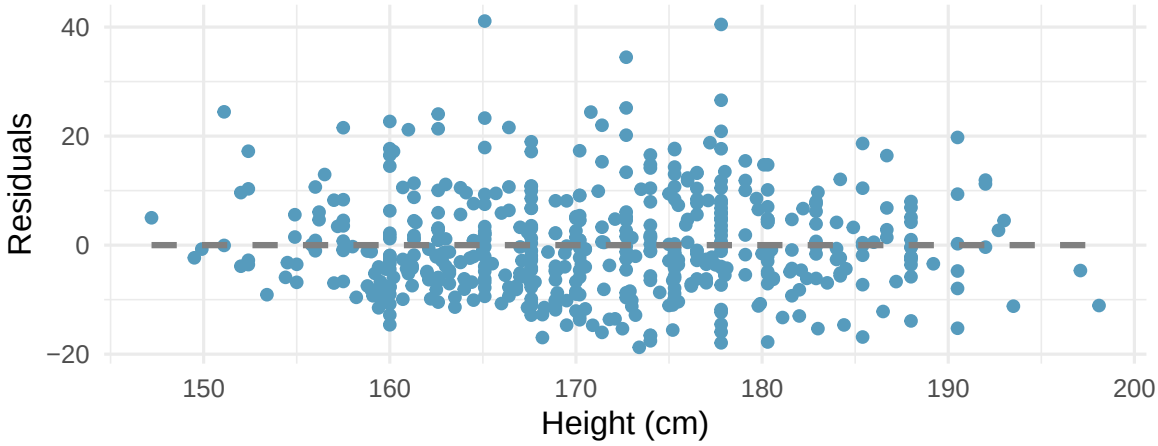
a. Describe the relationship between the percent of families who own their home and the percent of the population living in urban areas.
b. The outlier at the bottom right corner is District of Columbia, where 100% of the population is considered urban. What type of an outlier is this observation?



10. **Crawling babies, outliers.** A study conducted at the University of Denver investigated whether babies take longer to learn to crawl in cold months, when they are often bundled in clothes that restrict their movement, than in warmer months. The plot below shows the relationship between average crawling age of babies born in each month and the average temperature in the month when the babies are six months old. The plot reveals a potential outlying month when the average temperature is about 53F and average crawling age is about 28.5 weeks. Does this point have high leverage? Is it an influential point? (Benson 1993)



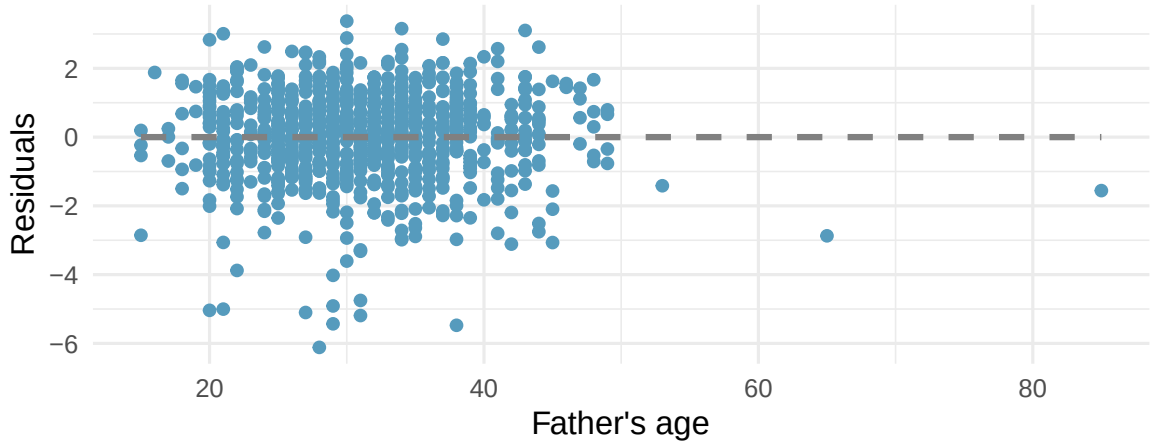
11. **Body measurements, conditions.** The scatterplot below shows the residuals (on the y-axis) from the linear model of weight vs. height from a dataset of body measurements from 507 physically active individuals. The x-axis is the height of the individuals, in cm. (Heinz et al. 2003)



a. For these data, R^2 is 51.84%. What is the value of the correlation coefficient? How can you tell if it is positive or negative? (Hint: you may need to look at a previous exercise.)

b. Examine the residual plot. What do you observe? Is a simple least squares fit appropriate for these data? Which of the LINE conditions are met or not met?

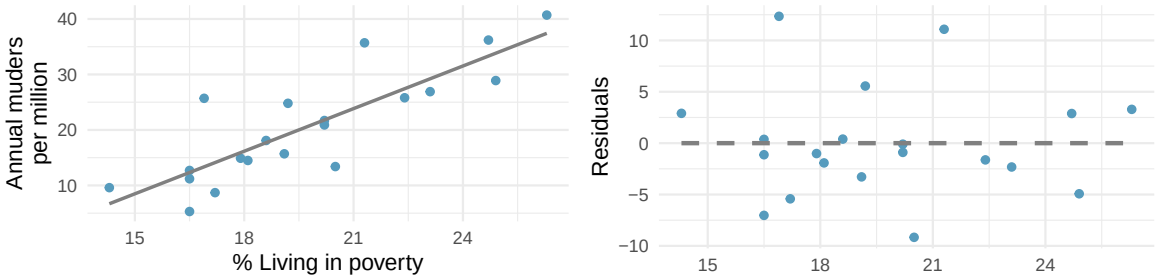
12. **Baby's weight and father's age, conditions.** The scatterplot below shows the residuals (on the y-axis) from the linear model of baby's weight (measured in pounds) vs. father's age for a random sample of babies. Father's age is on the x-axis. (ICPSR 2014)



a. For these data, R^2 is 0.09%. What is the value of the correlation coefficient? How can you tell if it is positive or negative? (Hint: you may need to look at a previous exercise.)

b. Examine the residual plot. What do you observe? Is a simple least squares fit appropriate for these data? Which of the LINE conditions are met or not met?

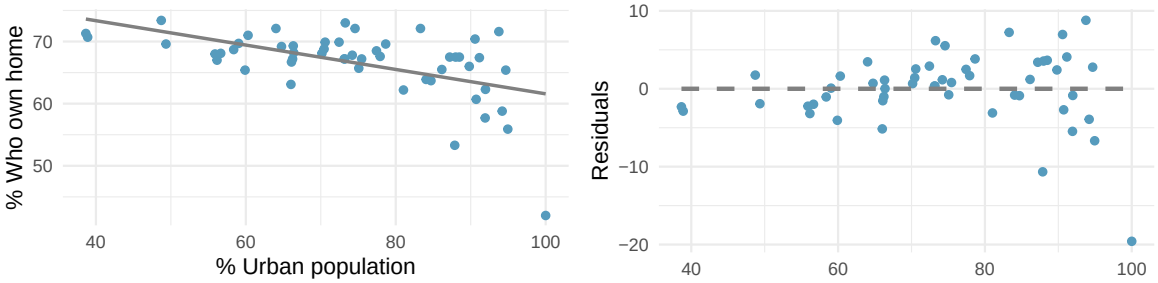
13. **Murders and poverty, conditions.** The scatterplot below shows the annual murders per million vs. percentage living in poverty in a random sample of 20 metropolitan areas. The second figure plots residuals on the y-axis and percent living in poverty on the x-axis.



a. For these data, R^2 is 70.56%. What is the value of the correlation coefficient? How can you tell if it is positive or negative?

b. Examine the residual plot. What do you observe? Is a simple least squares fit appropriate for the data? Which of the LINE conditions are met or not met?

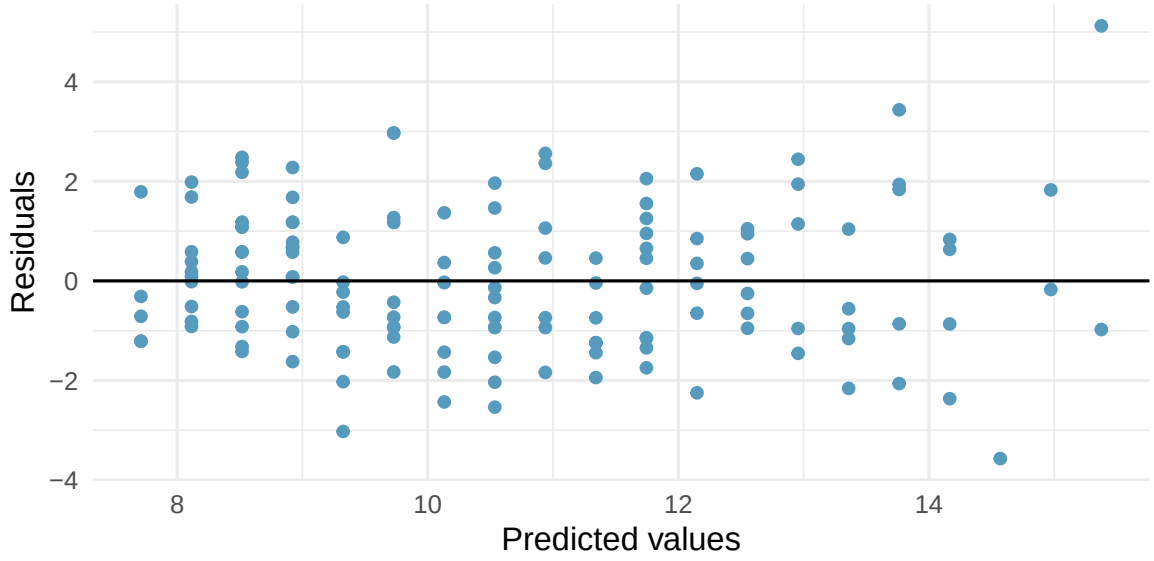
14. **Urban homeowners, conditions.** The scatterplot below shows the percent of families who own their home vs. the percent of the population living in urban areas. (US Census Bureau 2010) There are 52 observations, each corresponding to a state in the US. Puerto Rico and District of Columbia are also included. The second figure plots residuals on the y-axis and percent of the population living in urban areas on the x-axis.



a. For these data, R^2 is 29.16%. What is the value of the correlation coefficient? How can you tell if it is positive or negative?

b. Examine the residual plot. What do you observe? Is a simple least squares fit appropriate for the data? Which of the LINE conditions are met or not met?

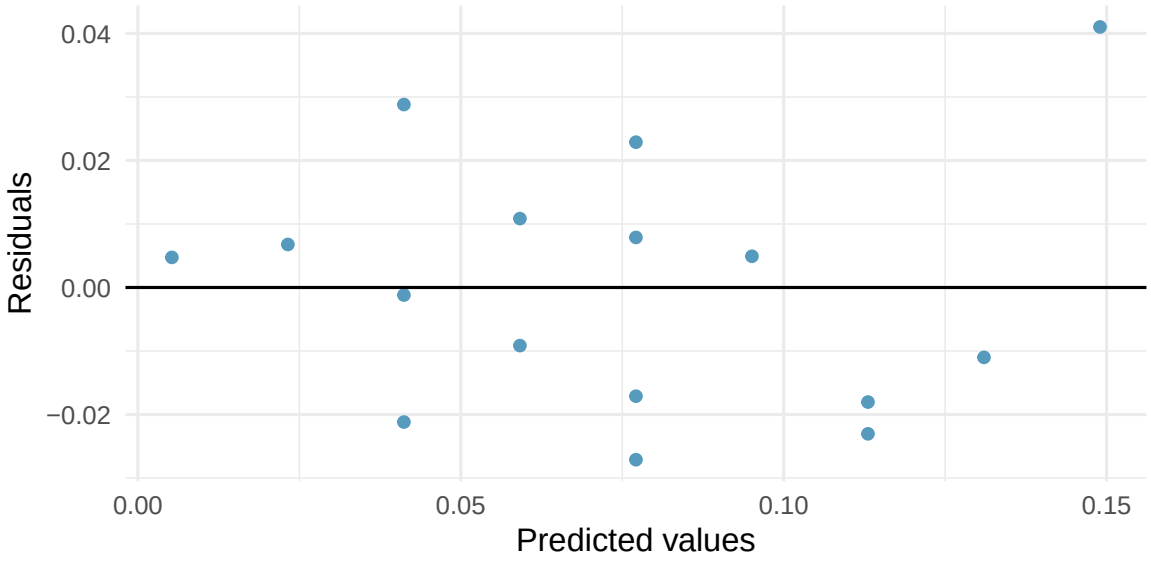
15. **I heart cats, LINE conditions.** Researchers collected data on heart and body weights of 144 domestic adult cats. The figure below shows the output of the predicted values and residuals generated from a linear model predicting heart weight (measured in grams) from body weight (measured in kilograms) of these cats.



a. Examine the residual plot. Notice that for the small predicted values the residuals have a smaller magnitude than the larger residuals seen with the larger predicted values. The change in magnitude of the residuals across the predicted values is an indication of violation of which LINE technical condition?

b. If the LINE condtion described in part (a) is violated, might it lead to an incorrect conclusion about the model (i.e., the least squares regression line itself), the inference of the model (i.e., the p-value associated with the least squares regression line), neither, or both? Explain your reasoning.

16. **Beer and blood alcohol content, LINE conditions.** The figure below shows the output of the predicted values and residuals generated from a linear model predicting the blood alcohol content (BAC) from number of cans of beer drunk by sixteen student volunteers at Ohio State University. (Malkevitch and Lesser 2008)



a. Examine the residual plot. Notice that it is difficult to identify any convincing patterns for or against violation of the LINE technical conditions. What is it about the residual plot that makes it difficult to assess the LINE technical conditions?

b. Is there anything about the residual plot which would make you hesitate about using the linear model for inference about all students? Is there anything about the experimental design of the study which would make you hesitate about using the linear model for inference about all students?

Datasets: `starbucks` (openintro) | `urban_owner` (usdata) | `babies_crawl` (openintro) | `bac` (openintro)

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Read residuals.** Open a regression dataset in the Regression Explorer and inspect the residual plot.
   a. What does a "good" residual plot look like? Sketch three features in words.
   b. Identify three *bad* patterns and name what each one tells you about the model.
   c. A student says: "If the residuals are small, the model is good." Why is the *size* of residuals not the main thing to look at?
   d. Suppose you see a fan-shaped residual plot (spread grows with x). Which regression assumption has failed? What does this do to the trust you can place in the reported t-test for the slope and its CI?

2. **R^2 in plain English.** A regression of campus weekly food spending on monthly income reports $R^2 = 0.41$.
   a. Translate "$R^2 = 0.41$" into one plain-English sentence about what fraction of variability in food spending is explained.

b. Does $R^2 = 0.41$ mean the regression is "weak" or "strong"? Why is this judgment context-dependent?
c. Adding a second predictor (e.g., number of dining-out trips) raises R^2 to 0.55. Is this *always* a sign that the new predictor matters? What does adding a *random* variable do to R^2?
d. A student concludes: "$R^2 = 0.41$ means 41% of food spending is *caused* by income." What's wrong with this?

3. **Extrapolation: don't trust the line outside the data.** A regression of city temperature on month of year (1–12) gives a fitted line over the year.

a. Why is using the line to predict the temperature *at month 25* not just unreliable, but conceptually wrong?
b. In a regression of test score on hours studied (in a sample where hours range 0–10), would you trust a prediction at 15 hours? At 5 hours? At -2 hours? Rank the trust.
c. Open any regression dataset in the Regression Explorer and note the range of x values. What would extrapolating the fitted line beyond that range tell you about the prediction — and what would it *not* tell you?
d. A real-estate analyst predicts the price of a 50000-square-foot mansion using a regression fit to homes 500–3500 sqft. Why is even a "narrow" prediction interval at $x = 50000$ deeply misleading?

4. **CI for the mean response vs. prediction interval for a new y.** These two intervals answer different questions at the same x_0.

a. In words, what does a **CI for the mean response at** x_0 estimate? Use a real-estate example to make it concrete.
b. In words, what does a **prediction interval for a new** y **at** x_0 estimate? Same example.
c. Which interval is **wider**, and why? (Hint: the prediction interval has two sources of uncertainty; the mean-response CI has only one.)
d. An analyst quotes the *mean-response CI* but speaks as if it were a *prediction interval* (e.g., "we're 95% sure the next house at 2000 sqft sells for between \$280k and \$320k"). What's wrong, and what is the right way to state each interval?

5. **Leverage and influential observations.** Open the Regression by Eye lab and generate a scatter with $n = 20$ near a moderate slope. Then mentally add a single extra point with very large x.

a. Define **leverage**. Why does an observation at extreme x have high leverage even before you look at its y?
b. Suppose your extra extreme-x point sits **on the fitted line** (it's high-leverage but consistent). Does removing it change the slope much? What does that tell you about high leverage in isolation?
c. Suppose instead the extra point sits **far off** the line — high leverage *and* high residual. Now does removing it change the slope? What's the buzzword for this kind of point?
d. In the Regression Explorer, find a real dataset and identify any potential high-leverage points by eye. What two questions should you ask before deciding whether to keep or remove them?

6. **Apply-unit capstone — write conclusions across every procedure in the book.** You have now met every inferential procedure this course covers: the one- and two-proportion z, the one-sample / two-sample / paired t, the chi-square goodness-of-fit and independence tests, one-way ANOVA, and slope inference. This capstone exercises the writing skill across all of them.

Open Conclusion Practice — Apply scope. The pool now includes chi-square scenarios in addition to the Calculate-unit scenarios you saw at the end of Ch 13. (*As of this printing, the practice*

tool does not yet include ANOVA or slope scenarios — see contract REQ-035 for planned additions. When those land, the same link picks them up automatically.)

a. Play at least **8 rounds** across a mix of procedures. Track which procedures give you the most trouble and which come naturally.
b. For at least one round where you scored *poorly*, re-read the scenario and identify the specific misinterpretation. Was it a p-value trap, a CI trap, a causation-vs-association slip, or a scope-of-inference issue?
c. Compare your performance now to the ch 13 Calculate-scope capstone (if you did that one). Have you internalized the reasoning skills better, or did the added chi-square scenarios trip you up? What does that tell you about your understanding of "one framework, many procedures"?
d. Finally: in one paragraph, write your own summary of the *inference framework* that unifies every procedure you learned this term. What are the four moves that appear in **every** test, regardless of the specific tool?

Chapter 22

ANOVA for Regression

In the linear regression chapter we tested a single hypothesis: is the population slope β_1 zero? That test used a t-statistic, and it answered a single question about a single coefficient. In this chapter we develop a *different* view of regression that scales naturally to models with many predictors: the **analysis of variance (ANOVA) table for regression**. This view partitions the total variability in the response into a piece the model explains and a piece it doesn't, and summarizes the resulting evidence for the model *as a whole* with an F-statistic. In simple regression the ANOVA F-test is algebraically equivalent to the t-test on β_1 (specifically, $F = t^2$), so it provides no new answers there — but the framework is the one that carries over to multiple regression, where it becomes essential.

See it in action. Open the Slope tool with ANOVA panel to display the ANOVA table for a simple regression fit (SS partition, F-statistic, model R^2), with the $F = t^2$ equivalence shown below the table. Pair with the Regression Explorer to see the fitted line and residuals side by side.

22.1 The two-question view of regression

When we fit a simple linear regression model $\hat{y} = b_0 + b_1 x$ to sample data, we can ask two related but distinct questions:

1. **Is the slope different from zero?** This is a *coefficient-level* question. In the linear regression chapter we answered it with a t-test on b_1.
2. **Is the model useful at all?** This is a *model-level* question. Does the fitted regression explain a discernible share of the variability in y, compared to using $\bar{y}$ as a "prediction" for every case?

In simple regression these two questions have the same answer, because a single-predictor model is "useful" exactly when its one slope coefficient is nonzero. When we add more predictors in the next chapter, the two questions come apart — some coefficients can be individually nonsignificant while the model as a whole is highly significant, or vice versa. The ANOVA-for-regression view is the tool we use to ask the *model-level* question.

22.2 Partitioning the total sum of squares

The **total variability** in the response is measured by

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

the sum of squared deviations of each observation from the sample mean. For any fitted regression line, we can split each deviation $y_i - \bar{y}$ into two pieces:

$$y_i - \bar{y} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{explained by the model}} + \underbrace{(y_i - \hat{y}_i)}_{\text{unexplained (the residual)}}.$$

Squaring and summing gives an identity that is arguably the single most important formula in this chapter:

The sum-of-squares decomposition for regression.

$$SST = SSR + SSE,$$

where

- $SST = \sum(y_i - \bar{y})^2$ is the **total sum of squares** (total variability in y),
- $SSR = \sum(\hat{y}_i - \bar{y})^2$ is the **regression sum of squares** (variability *explained* by the model),
- $SSE = \sum(y_i - \hat{y}_i)^2$ is the **error** or **residual sum of squares** (variability *not* explained).

The cross-product term that would appear in a naive expansion vanishes because the ordinary least-squares fit is orthogonal to the residuals by construction; you can take that as a promise for now.

This decomposition also gives an interpretation of R^2 that is the same one you already know but written in the new language:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

R^2 is the fraction of total variability that the regression model explains.

22.3 The ANOVA table for regression

Software organizes the sum-of-squares decomposition, degrees of freedom, and derived quantities into an **ANOVA table**:

Table 22.1: The ANOVA table for a simple linear regression fit. In simple regression the numerator degrees of freedom is always 1 (one slope estimated); the denominator is $n - 2$ (one intercept and one slope).

Source	df	SS	MS	F	p-value
Regression	1	SSR	$MSR = SSR/1$	$F = MSR/MSE$	(from $F_{1,\,n-2}$)
Residual	$n-2$	SSE	$MSE = SSE/(n-2)$		
Total	$n-1$	SST			

Two of the columns deserve extra comment.

- The **degrees of freedom** row is what you would guess: we lose two df fitting the two coefficients (b_0 and b_1), leaving $n - 2$ for the residual row; the regression row picks up the remaining 1 so that everything sums to the total's $n - 1$.
- MSE estimates the residual variance σ^2 around the true regression line. Its square root, $s = \sqrt{MSE}$, is the **residual standard error** that software reports on the same output.

22.4 The F-statistic and its distribution

The **F-statistic** for regression is a ratio of two variance estimates:

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}.$$

Under $H_0 : \beta_1 = 0$ (and the LINE conditions from the linear regression chapter), this statistic follows an F distribution with 1 numerator degree of freedom and $n - 2$ denominator degrees of freedom. When the true slope is zero, both MSR and MSE estimate the same noise variance σ^2, so F hovers near 1. When the slope is nonzero, the model captures real signal and MSR inflates relative to MSE, driving F above 1. So the test is always right-tailed: we compute a p-value as the area *to the right* of the observed F on the $F_{1,\,n-2}$ curve.

The **model-level F-test** for a simple linear regression tests

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_A : \beta_1 \neq 0$$

with test statistic $F = MSR/MSE$. Under H_0 and the LINE conditions, $F \sim F_{1,\,n-2}$. Small p-values (large F) provide evidence that the regression model captures real signal in y.

22.5 Worked example: possum head length on total length

The linear regression chapter fit a model of possum head length (mm) on total length (cm) using $n = 104$ possums. Suppose the software reports the following:

Quantity	Value
Slope estimate b_1	0.573
$SE(b_1)$	0.059
t-statistic for $H_0 : \beta_1 = 0$	9.66
Residual standard error s	2.60 mm
R^2	0.478

Fill in the ANOVA table.

Total sum of squares. With $n = 104$, the residual row has $df = n - 2 = 102$. The regression row has $df = 1$. The total row has $df = n - 1 = 103$.

The residual mean square is $MSE = s^2 = 2.60^2 = 6.76$, so

$$SSE = df_{\text{resid}} \times MSE = 102 \times 6.76 = 689.5.$$

$R^2 = 0.478$ means the model explains 47.8% of the total variability, so

$$SSR = \frac{R^2}{1 - R^2} \cdot SSE = \frac{0.478}{0.522} \times 689.5 \approx 631.$$

(Equivalently, $SST = SSE/(1 - R^2) \approx 689.5/0.522 \approx 1321$ and $SSR = SST - SSE \approx 631$.)

The regression mean square is $MSR = SSR/1 = 631$, and

$$F = \frac{MSR}{MSE} = \frac{631}{6.76} \approx 93.3.$$

Compare to $t^2 = 9.66^2 \approx 93.3$ — the two match to within rounding.

Under H_0 the null distribution is $F_{1,102}$, and $F = 93.3$ is far into the right tail; the p-value from software is essentially zero. The ANOVA F-test rejects the null of no linear relationship, in exact agreement with the slope t-test ($t = 9.66$) in the software output above.

22.6 $F = t^2$ in simple regression

The near-equality of F and t^2 in the worked example is not a coincidence.

In a simple linear regression, the ANOVA F-statistic and the slope t-statistic satisfy

$$F = t^2,$$

and the two tests give **identical** p-values. Rejecting $H_0 : \beta_1 = 0$ with the t-test is the same event as rejecting the model-level H_0 with the F-test.

This equivalence is a special case of a broader algebraic fact about F distributions: an F random variable with 1 numerator degree of freedom is the square of a t random variable with the corresponding denominator degrees of freedom. So in simple regression the two tests are literally the same test dressed in different notation.

Why bother introducing the ANOVA framework at all, then? Because in **multiple regression** — with two or more predictors — the F-test tests whether *any* of the slopes are nonzero. There is no single t-statistic that answers that question; each predictor gets its own t, and the model-level F combines them. The ANOVA-for-regression view is the correct generalization; the coincidence of $F = t^2$ in the one-predictor case simply reflects that the two views have to agree when there is only one slope to test.

22.7 Bridge to multiple regression

The next chapter introduces regression models with **more than one** predictor. The ANOVA framework extends naturally.

- The **regression sum of squares** picks up the variability jointly explained by all the predictors together; its degrees of freedom equal the number of predictors k.
- The **residual sum of squares** measures the variability the multi-predictor model still cannot explain; its degrees of freedom are $n - k - 1$ (one lost per slope plus one for the intercept).
- The **overall F-statistic** tests $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$, i.e. "none of the predictors help." Under H_0 it follows an $F_{k,\, n-k-1}$ distribution.
- The **coefficient-level t-tests** answer individual questions about each predictor "controlling for the others." A predictor with a discernible t-test contributes above and beyond the other predictors already in the model.

In simple regression there is only one t-test and one F-test and they agree; in multiple regression there are $k + 1$ tests and they can disagree in informative ways. The rest of that story is Chapter 23.

22.8 Chapter review

22.8.1 Summary

This chapter introduced the ANOVA-for-regression framework. The total sum of squares SST decomposes as $SST = SSR + SSE$, splitting the response's variability into a piece the model explains and a piece it does not. Organizing this decomposition alongside degrees of freedom, mean squares, and an F-statistic gives the **ANOVA table for regression**. The F-statistic tests the model as a whole; in simple regression it satisfies $F = t^2$ and gives the same p-value as the slope t-test, but the framework generalizes to multiple regression where the two views come apart. Along the way we reinterpreted $R^2 = SSR/SST$ as the fraction of total variability the model explains.

22.8.2 Terms

SST, SSR, SSE, sum-of-squares decomposition, regression mean square MSR, residual mean square MSE, ANOVA table for regression, model-level F-test, F distribution, residual standard error, $F = t^2$ equivalence in simple regression.

22.8.3 Attribution

This chapter's ANOVA-for-regression framing — SS partition, F-statistic, R^2 through the ANOVA lens, and residual standard error — is presented here as a unified treatment. The two contemporary CC BY-SA OER intro-stats texts (IMS 2e's `inf-model-slr.qmd` and OpenIntro Statistics's regression chapters) both use the coefficient-level t-test route to slope inference without a formal ANOVA table; the exposition here is authored fresh rather than remixed from either. The $F = t^2$ equivalence and the bridge section to multiple regression are original to this chapter.

22.9 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Reading the ANOVA table.** A regression of highway miles-per-gallon on curb weight for $n = 60$ vehicles produces the following (abbreviated) ANOVA table:

Source	df	SS	MS	F
Regression		442		
Residual		128		
Total				

   (a) Fill in the missing degrees of freedom, mean squares, total sum of squares, and F-statistic.
   (b) Compute R^2.
   (c) State the null and alternative hypotheses that the F-test evaluates. Given the value of F you computed, is the p-value small or large? (You may reason from the size of F alone.)

2. **Filling gaps from R^2 and s.** A regression of test scores on hours studied for $n = 25$ students reports residual standard error $s = 4.20$ (points) and $R^2 = 0.36$.

   (a) Compute MSE and SSE.
   (b) Compute SST and SSR.
   (c) Report the ANOVA F-statistic.

3. **The $F = t^2$ equivalence.** A software output reports $t = 3.1$ for the slope with 47 degrees of freedom, and $F = 9.61$ for the model.
   (a) Show numerically that $F = t^2$ for this fit.
   (b) Explain in one sentence why the two tests must give the same p-value.
4. **Interpreting R^2 using sums of squares.** Suppose a fitted regression has $SSR = 200$ and $SSE = 800$.
   (a) Compute R^2.
   (b) In plain English, what does the value of R^2 tell you about how useful the regression is at explaining y?
   (c) Would you feel confident using this model for prediction? What other information would you want to see before deciding?
5. **When the F-test does *not* reject.** A study of $n = 40$ light bulbs regresses lifetime on bulb voltage. The ANOVA reports $F = 0.34$ with 1 and 38 degrees of freedom.
   (a) Estimate the p-value roughly (you may use the "F near 1 means no evidence" rule).
   (b) What conclusion does this suggest about the linear relationship between voltage and lifetime in this sample?
   (c) Does a non-rejecting F-test mean there is no relationship at all between voltage and lifetime? What are two alternative explanations?

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Frame the question.** A researcher fits a simple linear regression of a student's final exam score on hours studied.
   a. Write the null and alternative hypotheses that the **model-level F-test** evaluates. Write them again for the **slope t-test**.
   b. Why are these the *same* hypothesis in a simple regression? What would need to be true for them to differ?
   c. Suppose the F-statistic comes out to 12.25 with 1 and 48 degrees of freedom. What t-statistic on 48 df would give the same p-value?
2. **Open the ANOVA panel.** Open the Slope tool with ANOVA panel for the possum head-length regression.
   a. Report the sums of squares SSR, SSE, SST and verify that $SSR + SSE = SST$.
   b. Report the degrees of freedom for each row. Explain why $df_{\text{regression}} = 1$ and $df_{\text{residual}} = n - 2$.
   c. Report the F-statistic and the t-statistic for the slope, and confirm $F = t^2$.
   d. Report R^2. Verify from the ANOVA table that $R^2 = SSR/SST$.
3. **Read R^2 correctly.** A fitted regression reports $R^2 = 0.72$.
   a. Complete this sentence: "About ____ % of the variability in the response is explained by ____." Which value goes in each blank?
   b. A student concludes: "$R^2 = 0.72$ means my model is 72% accurate." Why is this wrong, and what's the correct statement?
   c. What does $R^2 = 0.72$ *not* tell you about the model? Give three specific things.
4. **When F rejects and t agrees.** Open `inference/slope/?anova=true` on any regression dataset with a clear linear trend.

a. Report F, its degrees of freedom, and the p-value.
b. Now switch datasets to one you expect to have *no* linear relationship. Report F and the p-value. Is F close to 1?
c. Explain in one sentence why "F close to 1" is the "no linear signal" case.

5. **The multiple-regression preview.** In simple regression, the model-level F-test and the slope t-test *must* agree. In multiple regression, they can diverge.

   a. Imagine a multiple regression with 3 predictors where the overall F-test rejects at $\alpha = 0.05$ but *no* individual predictor's t-test does. What could explain this? (Hint: think about *joint* vs. *individual* contributions.)
   b. Imagine the reverse: no individual t-test rejects, and the F-test doesn't either. What does this suggest about the model?
   c. Imagine one predictor's t-test rejects while the model-level F-test does not. Which conclusion should you trust?

Chapter 23

Multiple Regression

The regression models in earlier chapters used a single predictor x to explain a response y. Most real datasets carry many potential predictors: the interest rate a lender charges depends on a borrower's income *and* the loan term *and* their credit history; a penguin's body mass depends on flipper length *and* bill length *and* species. **Multiple linear regression** extends the simple regression framework to accommodate two or more predictors at once. This chapter follows the structure of the IMS 2e chapter *Inference for linear regression with multiple predictors* (`inf-model-mlr.qmd`, CC BY-SA 3.0) — adopting IMS's running loans case study, its coin-dish multicollinearity example, and its conditional hypothesis-testing framing — while writing the exposition anew for the simulation-first sequence. Coverage of cross-validation, formal model-selection procedures, adjusted R^2, and the variance inflation factor is intentionally kept light or moved to Extended Content in this chapter.

See it in action. Open the Multiple Regression tool to pick a response and predictors, read the coefficient table, model F-test, R^2/adjusted R^2, and diagnostics (residuals, Cook's distance). The base view shows a **pairwise predictor scatterplot matrix** for spotting collinearity by eye; append `?expert=true` for the **VIF** column (Extended Content). Try possum with two predictors to reproduce this chapter's worked example.

23.1 From one predictor to many

A single-predictor regression is a starting point, not an ending point. In the real world, a response variable is almost never driven by just one thing. When we ignore other variables that also influence the response, two things can go wrong.

Confounding can bias the slope estimate. Suppose we regress interest rate on debt-to-income ratio alone. If borrowers with higher debt-to-income also tended to have longer loan terms and more credit checks, the fitted slope would absorb part of the effects of those variables too. Our estimate of "how much does an extra unit of debt-to-income cost, for a comparable borrower?" would then be contaminated by the unmeasured differences in loan term and credit history.

Predictive accuracy suffers. Even when we don't care about causal interpretation, ignoring available predictors leaves signal on the table. A model with three good predictors typically predicts future observations more accurately than a model with one.

Multiple regression solves the first problem by fitting all predictors together in a single model, so each coefficient represents the effect of that predictor *while holding the others fixed*. And it addresses the second by explaining more of the total variability with the additional predictors.

23.2 The multiple regression model

The **multiple linear regression model** with k predictors is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

where β_0 is the population intercept, $\beta_1, \ldots, \beta_k$ are the population slopes for the k predictors, and ε is a random error term with mean 0 and constant variance σ^2.

The fitted model estimated from a sample is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k,$$

with $b_0, b_1, \ldots, b_k$ chosen to minimize the sum of squared residuals $\sum (y_i - \hat{y}_i)^2$ — the same least-squares criterion used in simple regression.

Geometrically, a single-predictor regression fits a *line* in a two-dimensional scatterplot; a two-predictor regression fits a *plane* in a three-dimensional cloud. With three or more predictors we lose the ability to draw the picture, but the mathematics generalizes without incident: we are always fitting the flat surface that comes closest to the observed points in the sum-of-squared-residuals sense.

23.3 Reading multiple-regression output: the loans case study

Predicting interest rates. IMS's running case study uses `loans_full_schema` from the **openintro** R package — a curated sample of loans made through a peer-to-peer lending platform. The response is the annual **interest rate** charged; predictors include the borrower's **debt-to-income** ratio, the loan **term** (36 or 60 months), and the number of credit **inquiries** in the past 12 months (`credit_checks`).

Fitting the model

$$E[\texttt{interest\_rate}] = \beta_0 + \beta_1 \cdot \texttt{debt\_to\_income} + \beta_2 \cdot \texttt{term} + \beta_3 \cdot \texttt{credit\_checks}$$

in software gives the following coefficient table (numbers from IMS 2e):

Predictor	Estimate	Std. Error	t	p-value
(Intercept)	4.31	0.20	22.1	< 0.0001
debt_to_income	0.041	0.003	13.3	< 0.0001
term	0.16	0.004	37.9	< 0.0001
credit_checks	0.25	0.019	12.8	< 0.0001

The fitted equation is

$$\widehat{\texttt{interest\_rate}} = 4.31 + 0.041 \cdot \texttt{debt\_to\_income} + 0.16 \cdot \texttt{term} + 0.25 \cdot \texttt{credit\_checks}.$$

Interpret each coefficient in context.

- **Intercept.** A hypothetical borrower with debt-to-income 0, term 0, and 0 credit checks would receive an estimated mean interest rate of 4.31%. As is often the case, this hypothetical borrower doesn't exist (a loan can't have a term of 0), so the intercept mostly serves to position the fitted plane, not as a substantively meaningful prediction.
- **debt_to_income.** Holding term and credit checks fixed, each additional unit of debt-to-income ratio is associated with a 0.041 percentage-point increase in the mean interest rate.
- **term.** Holding debt-to-income and credit checks fixed, each additional month of loan term is associated with a 0.16 percentage-point increase in the mean interest rate. (Since term in this dataset takes only the values 36 or 60, the practical interpretation is that the 60-month loans carry a rate about $24 \times 0.16 \approx 3.8$ percentage points higher than the 36-month loans, on average.)
- **credit_checks.** Holding debt-to-income and term fixed, each additional credit inquiry in the past 12 months is associated with a 0.25 percentage-point increase in the mean interest rate.

Each of the four t-statistics has a p-value below 0.0001, so each predictor contributes discernibly *above and beyond* the others.

23.4 Interpreting coefficients "holding others constant"

The single most important interpretive habit in multiple regression is the phrase **"holding the other predictors constant."** It appears in every coefficient interpretation, and dropping it turns a correct sentence into a wrong one.

To see why, consider the loans model above. A simple regression on `debt_to_income` alone has no way to separate debt-to-income from anything that travels with it. In these data, borrowers with higher debt-to-income are only slightly more likely to take out longer-term loans and to have more credit checks (correlations of about 0.05 and 0.03), so the simple-regression slope, 0.047, is only a little larger than the multiple-regression slope. When predictors are more strongly related to each other, the gap between the simple and multiple slopes can be much larger.

The multiple-regression slope $b_1 = 0.041$ isolates the debt-to-income effect from the term and credit-check effects. That is why it can be interpreted as "for each additional unit of debt-to-income, holding term and credit checks fixed." The "holding fixed" is not a philosophical qualifier; it is the *only* way the coefficient makes sense.

Every coefficient interpretation in multiple regression should include the phrase **"holding the other predictors constant"** (or an equivalent like "controlling for" or "at fixed values of"). Interpreting b_j *without* that qualifier is one of the most common errors students make with multiple regression output.

23.5 Testing individual coefficients

Each row of the coefficient table also reports a t-statistic and p-value. These test the null hypothesis that a single population slope is zero — but now the hypothesis is *conditioned* on the other predictors being in the model.

Conditional slope hypotheses. In a multiple regression with k predictors, the individual t-test for predictor x_j tests

$$H_0 : \beta_j = 0, \quad \text{given the other } k - 1 \text{ predictors are included in the model.}$$

A small p-value provides evidence that x_j contributes to explaining the response *above and beyond* what the other predictors already explain.

Rejecting $H_0 : \beta_j = 0$ tells us that dropping x_j from the model would meaningfully worsen the fit. Failing to reject does *not* tell us that x_j is unrelated to y — it might be strongly related, but the *unique* piece it contributes above what the other predictors already capture is small. This distinction matters most when predictors are correlated with each other, which is the topic of the multicollinearity section.

23.6 ANOVA for multiple regression

The ANOVA-for-regression framework introduced in the ANOVA-for-Regression chapter carries over to multiple predictors with only two changes.

For a multiple regression with k predictors and n observations, the ANOVA table is:

Source	df	SS	MS	F
Regression	k	SSR	$MSR = SSR/k$	$F = MSR/MSE$
Residual	$n-k-1$	SSE	$MSE = SSE/(n-k-1)$	
Total	$n-1$	SST		

Under $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (that *no* predictor contributes) and the LINE conditions, the F-statistic follows an $F_{k,\,n-k-1}$ distribution. A small p-value indicates that the model as a whole explains a discernible share of the variability in y; it does not tell us *which* predictors are responsible. The coefficient-level t-tests answer the "which one" question — with the conditional caveat above.

The interpretation of $R^2 = SSR/SST$ as "the fraction of total variability the model explains" is unchanged from simple regression. The important new fact is that adding *any* predictor to the model — even a completely useless one — can only *increase* R^2, never decrease it. This is what motivates the Extended Content on adjusted R^2 below.

Extended Content: adjusted R^2.
Because R^2 never decreases when we add predictors, comparing candidate models with different numbers of predictors purely by R^2 is unfair to the smaller model. **Adjusted** R^2 corrects for that unfairness:

$$R^2_{\text{adj}} = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1}.$$

Adding a useless predictor eats up one degree of freedom in the denominator without meaningfully reducing SSE, so R^2_{adj} *decreases* rather than increases. Adjusted R^2 is the standard summary to compare models with different numbers of predictors; it is beyond a typical introductory track but is retained here for readers who encounter it in later coursework.

23.7 Multicollinearity: the coin-dish story

Two predictors in a regression exhibit **multicollinearity** when they are strongly linearly related to *each other*. This is not a problem for prediction, but it is a serious problem for interpreting individual coefficients: when two predictors carry nearly the same information, the fit cannot tell them apart, and the individual t-tests may fail to reject even when the joint effect of the pair is strong.

The classic teaching example (from IMS 2e, ultimately due to Jeff Witmer at Oberlin) is a coin-dish scenario. Suppose we sample 26 students and ask each to report

- the **total amount of money** in their coin dish (dollars),
- the **total number of coins** in the dish,
- the **number of low coins** (coins that are not quarters).

Two simple regressions. Fitting the two simple regressions separately gives (numbers from IMS 2e):

- $\widehat{\text{total\_amount}} = 0.55 + 0.13 \cdot \text{number\_of\_coins}$ — slope $b_1 = 0.13$, p-value very small.
- $\widehat{\text{total\_amount}} = 2.28 + 0.02 \cdot \text{number\_of\_low\_coins}$ — slope $b_1 = 0.02$, p-value large (no discernible relationship).

Interpret the two slopes and, briefly, why they take the values they do.

For every additional coin in the dish, the total amount goes up by about \$0.13 on average — roughly the average value of a US coin, as we'd hope. For every additional *low* coin (a penny, nickel, or dime), the total goes up by only about \$0.02 on average — which sounds reasonable but is barely discernible from zero, because adding a low coin only adds pennies, nickels, or dimes.

Now fit the model with *both* predictors together. IMS reports

$$\widehat{\text{total\_amount}} = 0.80 + 0.21 \cdot \text{number\_of\_coins} - 0.16 \cdot \text{number\_of\_low\_coins},$$

both coefficients discernible.

Explain what happened. Why is the coefficient on `number_of_low_coins` *negative* in the multi-predictor model when it was slightly *positive* on its own?

The reason is the "holding others constant" language. In the multi-predictor model,

- $b_1 = 0.21$ says: for each additional coin *while the number of low coins stays constant*, the total goes up by 21 cents. Increasing "total coins" while keeping "low coins" fixed means adding a *quarter* — which adds 25 cents on average. So the coefficient makes sense.
- $b_2 = -0.16$ says: for each additional low coin *while the total number of coins stays constant*, the total goes *down* by 16 cents. Swapping a quarter (25 cents) for a low coin (a few cents) drops the total by roughly that amount. So this coefficient also makes sense.

The two individual slopes are answering *different* questions because they condition on the other variable in different ways. When predictors are strongly related to each other, the multi-predictor coefficients can look very different from the simple-regression slopes — and *both* interpretations are correct, as long as we honor the "holding others constant" qualifier.

A telltale sign of multicollinearity is a predictor whose relationship to the response looks strong on a scatterplot yet has an individual t-test that fails to reject. The predictor may be redundant with another predictor already in the model, contributing no additional information once the correlated predictor is accounted for.

The most direct diagnostic is to **plot the predictors against each other**. A strong linear pattern between two predictors is evidence that they carry overlapping information. Common responses to strong pairwise associations: **(i) drop one of the collinear predictors**, **(ii) combine them into a single summary variable** (an average or sum, when units allow), or **(iii) accept that individual coefficients cannot be separately interpreted but that the model still predicts well as a whole**.

Extended Content: the variance inflation factor.
Beyond scatterplots, statisticians quantify the severity of multicollinearity with a single number per predictor: the **variance inflation factor (VIF)**. For predictor x_j,

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the R^2 of a regression of x_j on all the *other* predictors. A predictor uncorrelated with the others has $R_j^2 = 0$ and $VIF_j = 1$; a predictor perfectly explained by the others has $R_j^2 = 1$ and infinite VIF. **Rules of thumb** (not sharp thresholds): $VIF < 5$ is fine; $5 \leq VIF < 10$ warrants caution; $VIF \geq 10$ means individual coefficients are unreliable.
VIF is retained here for readers, instructors, and later coursework that want a formal metric; a typical introductory sequence detects multicollinearity through the scatterplot approach above and does not require the VIF machinery.

23.8 Categorical predictors: a brief note

The models above use numeric predictors only. Multiple regression also accommodates **categorical predictors** by encoding them as **indicator variables** (also called dummy variables). For a two-level categorical variable (say, `verified_income` = yes or no), we create a numeric variable that takes the value 1 when the category is "yes" and 0 otherwise; then this indicator enters the regression like any other numeric predictor. The coefficient on the indicator represents the difference in mean response between the two categories, holding the other predictors fixed. For categorical variables with three or more levels, the same idea applies but requires one indicator variable per non-baseline category.

A full treatment of indicator variables — including baseline choice, interpretation, and interaction terms — is beyond the scope of this chapter and belongs in a follow-on course. IMS 2e's `inf-model-mlr.qmd` chapter covers the extension in detail if you want the deeper version.

23.9 Diagnostics for multiple regression

The LINE conditions from the linear regression chapter carry over unchanged to multiple regression, but the checks look slightly different because we no longer have a single predictor's scatterplot to eyeball.

Linearity. Plot each predictor against the residuals. A curved pattern in one plot suggests that predictor's relationship with the response is not linear at fixed values of the others.

Independence. Same as in simple regression: the design of the study earns this assumption. Random sampling or random assignment protects independence.

Normality of residuals. A residual histogram or Q-Q plot works the same as in simple regression.

Equal variance. The residual-versus-fitted plot is still the go-to diagnostic. A fan shape suggests the residual spread grows with $\hat{y}$; if so, coefficient standard errors are unreliable and the reported CIs are too narrow.

Add one MLR-specific check:

Multicollinearity. Inspect scatterplots of each pair of predictors. Strong pairwise associations flag pairs that carry overlapping information about the response. (Extended-content readers can also inspect the variance inflation factor for a formal quantitative check.)

23.10 Chapter review

23.10.1 Summary

Multiple regression extends the simple regression model to two or more predictors, fit by least squares. The interpretation of each coefficient becomes "the mean change in y per unit change in x_j *holding the other predictors fixed*" — a qualifier that separates multiple regression's slope from the simple-regression slope that would absorb the other predictors' effects. Software reports coefficient estimates alongside a coefficient table, an ANOVA table with a model-level F-test, and the goodness-of-fit statistic R^2. Individual t-tests for each coefficient are *conditional* on the other predictors being in the model. Multicollinearity — overlapping information between predictors — is a new concern that only arises with multiple predictors; the standard check is a scatterplot of each pair of predictors, and the coin-dish example illustrates why interpretations of "simple" and "multiple" slopes can look wildly different when predictors are correlated. Diagnostics for the LINE conditions carry over from simple regression. Categorical predictors extend the framework through indicator variables and are covered in follow-on courses.

23.10.2 Terms

Multiple linear regression, predictor, response, coefficient table, conditional hypothesis test, "holding others constant," model-level F-test, multicollinearity, indicator variable, residual-versus-fitted plot, extrapolation in multi-dimensional predictor space. (*Extended-content only: adjusted R^2, variance inflation factor / VIF.*)

23.10.3 Attribution

This chapter is a remix of the IMS 2e chapter *Inference for linear regression with multiple predictors* (`inf-model-mlr.qmd`, Cetinkaya-Rundel and Hardin, CC BY-SA 3.0). The loans case study, the coin-dish multicollinearity example, and the conditional hypothesis-testing framing are IMS's; the exposition is rewritten. IMS's sections on cross-validation, formal model-selection procedures, adjusted R^2, and VIF are omitted or moved to Extended Content. Readers who want IMS's cross-validation treatment or the full indicator-variable extension should consult the IMS source directly.

23.11 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

Note. Exercises 3 (VIF interpretation) and 6 (adjusted R^2) draw on the **Extended Content** callouts in the chapter. Skip them if you are following the base track.

1. **Reading the loans coefficient table.** Recall the loans model from the chapter, predicting interest rate from debt-to-income, loan term (in months), and credit checks in the past 12 months:

$$\widehat{\texttt{interest\_rate}} = 4.31 + 0.041 \cdot \texttt{debt\_to\_income} + 0.16 \cdot \texttt{term} + 0.25 \cdot \texttt{credit\_checks}.$$

   (a) Predict the interest rate for a borrower with debt-to-income 15, a 36-month loan, and 2 credit checks in the last year.
   (b) Interpret the coefficient on `credit_checks` in context. Include the phrase "holding the other predictors constant."

(c) The individual t-test on `credit_checks` has a very small p-value. State the null hypothesis this test evaluates, in symbols *and* in words, including the "given other predictors are in the model" qualifier.

2. **Simple vs. multiple regression slope.** Suppose the *simple* regression of interest rate on `credit_checks` alone gives a slope of 0.28 — slightly larger than the multi-predictor slope of 0.25.

   (a) Why is the simple-regression slope larger?
   (b) Which slope answers the question "for two borrowers with the same debt-to-income and same loan term, how much does one extra credit check cost the borrower with more inquiries?"
   (c) Which slope answers the question "for two random borrowers differing only in credit-check count (with all their other characteristics also differing on average), how different are their expected interest rates?"

3. **VIF interpretation (Extended Content).** A multiple regression on employee salary reports the following VIFs:

Predictor	VIF
Age	1.4
Years of experience	8.7
Years at company	9.1
Job level	1.2

   (a) Which two predictors are causing concern about multicollinearity, and why?
   (b) Explain in one sentence why high VIF makes it hard to interpret those two coefficients individually.
   (c) Name two possible responses to the high VIFs.

4. **The coin-dish story.** Recall the coin-dish example from the chapter. The two simple regressions give slopes of 0.13 (on total number of coins) and 0.02 (on number of low coins). The multiple regression with both predictors gives slopes of 0.21 (coins) and -0.16 (low coins).

   (a) In the multi-predictor model, the slope on `number_of_low_coins` is *negative*. Explain in one sentence why this makes sense in context, using the "holding the other predictor fixed" framing.
   (b) A student says: "The simple regression shows low coins have a *positive* effect on total amount, but the multiple regression flips it to negative. One of them must be wrong." Push back on this.
   (c) When two predictors are strongly related to each other, the "same predictor" can appear to have a positive marginal effect and a negative conditional effect. What does this tell you about the danger of interpreting single-predictor slopes as if they were "the effect" of the predictor?

5. **When to add a predictor.** A researcher must choose whether to add a fifth predictor to a four-predictor model of house prices. The new predictor's individual t-test has p-value 0.28, and a scatterplot shows a moderately strong linear relationship between the new predictor and one of the four already in the model.

   (a) Should the researcher add the predictor purely based on the statistical evidence?
   (b) Suppose the researcher has a strong subject-matter reason to include the predictor even though its p-value is large. Is that a valid reason to include it?
   (c) The residual standard error (root MSE) after adding the predictor is essentially unchanged. What does this tell you about the predictor's contribution?

6. **Adjusted R^2 vs. R^2 (Extended Content).** Two candidate models are fit to the same data ($n = 60$):

   - Model A (3 predictors): $R^2 = 0.62$, adjusted $R^2 = 0.60$.
   - Model B (5 predictors): $R^2 = 0.65$, adjusted $R^2 = 0.59$.

   (a) Which model has more explained variability by raw R^2?
   (b) Which model is preferred by adjusted R^2? Why do they disagree?
   (c) Argue in one sentence why adjusted R^2 is the more honest comparison when the models have different numbers of predictors.

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

Note. Exercises 4 and 5 optionally use the **Extended Content** variance inflation factor material, available via **expert mode** on the Multiple Regression tool. The reasoning still works if you substitute "scatterplot showing a strong relationship between predictors" for high VIF and "no strong relationship on the predictor scatterplot" for low VIF.

1. **Frame the loans question.** A lender wants to build a model to predict a borrower's interest rate from `debt_to_income`, `term`, and `credit_checks` — the same three predictors the chapter uses.

   a. Write the multiple regression model in the standard form $y = \beta_0 + \beta_1 x_1 + \cdots$ with all variables clearly labeled.
   b. State the null and alternative hypotheses for the **model-level** F-test in words.
   c. State the null hypothesis for the **individual** t**-test** on `debt_to_income` — including the "given other predictors are in the model" qualifier. Why is this a different question from the one the F-test asks?

2. **Choose predictors thoughtfully.** For each scenario, decide whether to include the new predictor in the model. Options: **include**, **exclude**, or **it depends** (and specify the deciding factor).

   a. The new predictor's individual t-test has p = 0.03; its scatterplot vs. existing predictors shows no strong relationships; residual standard error drops noticeably after adding it.
   b. The new predictor's t-test has p = 0.42; residual standard error is essentially unchanged.
   c. The new predictor's t-test has p < 0.001; its scatterplot vs. an existing predictor shows a very strong linear relationship (or, equivalently, its VIF is 14.2).
   d. The new predictor has a strong scientific reason to be in the model (a well-established causal driver), t-test p = 0.20, no strong scatterplot associations with the other predictors.

3. **Read the coefficient table (possum data).** Open the Multiple Regression tool loaded with the possum data. Select **head length** as the response and both **total length** and **skull width** as predictors.

   a. Write the fitted equation with numeric coefficients.
   b. Interpret each slope in context, using the phrase "holding the other constant."
   c. Compare the total-length slope in the two-predictor model to the total-length slope from the *simple* regression in the linear-regression chapter. Are they the same? If they differ, what does that tell you about what the simple slope was actually measuring?
   d. Report the model-level F-statistic and its p-value. Does the model contribute discernibly?

4. **Multicollinearity via scatterplot (Extended: VIF).** Add a third predictor (say, **tail length**) to the model in Exercise 3.

   a. Open the **pairwise scatterplot matrix** of the three predictors. Are any pair of predictors visibly related? Which pair?

b. Report the coefficient for total length and its standard error. How does the standard error compare to the two-predictor version? What does the change reflect?
c. Optional (Extended Content, requires expert mode): Report the VIFs for all three predictors. Do the VIF values agree with what the pairwise scatterplot showed?
d. A student says: "adjusted R^2 went up, so the third predictor is clearly worth including." Is that always sufficient justification? What else should you check before finalizing the model?

5. **When F and t disagree.** Use the Multiple Regression tool to construct or find a dataset where the model-level F-test rejects at $\alpha = 0.05$ but *none* of the individual t-tests do.

a. Report the F-statistic, its p-value, and the p-values of the individual t-tests. What does this pattern suggest?
b. Inspect the pairwise scatterplot matrix of the predictors. Are any pairs strongly related? (*Optional Extended Content: report the VIFs.*)
c. What does "the model as a whole works, but no individual predictor is discernible" *mean* for the researcher trying to use this model?
d. A student says: "If no individual t-test rejects, we should throw out the model." Push back on that argument in one sentence.

Apply — Unit Review

Unit 4 extends inference to richer data: **categorical associations**, comparing **many groups**, and **relationships between numerical variables**. The procedures are new, but the habit is the same — read the variables (how many, what type, how many groups) and let that pick the method. This review drills exactly that match.

The big picture

Three families of questions, each tied to a data shape:

- **Categorical data → chi-square.** Does one categorical variable match an expected distribution (goodness of fit)? Are two categorical variables associated (test of independence)?
- **One numerical outcome across several groups → ANOVA.** Do three or more group means differ (ANOVA)? If so, *which* pairs differ (multiple comparisons)?
- **Two numerical variables → correlation and regression.** How strong and what direction is the linear association (correlation)? What line predicts one from the other, and is the slope real (linear regression)? How well does the model fit, and how far can you trust it (prediction & model fit)?

Which method?

Read the variables; the row names the method:

Variables in the question	Method	Statistic
One categorical vs. a claimed distribution	chi-square **goodness of fit**	χ^2
Two categorical (associated?)	chi-square **test of independence**	χ^2
One numerical outcome across **3+ groups**	**ANOVA** (then post-hoc comparisons)	**F**
Two numerical (strength/direction)	**correlation**	r
Two numerical (predict / is the slope real?)	**linear regression**	t on the slope

ANOVA answers "are *any* of the group means different?" — not which ones. A significant F says at least one pair differs; you need **multiple-comparison** procedures (which guard against the inflated error of many pairwise tests) to say *where* the difference is.

A quick decision guide

1. **What types are the variables?**
   - **All categorical** → chi-square. *One* variable vs. an expected split → **goodness of fit**; *two* variables → **independence**.
   - **A numerical outcome split by a categorical group → two means? use Unit 3 t; three or more groups? → ANOVA**.
   - **Two numerical variables → correlation** (how strong) and/or **regression** (predict, test the slope).
2. **Check the method's conditions** (e.g., chi-square expected counts; regression's linear, roughly-constant-spread residuals).

Key ideas to carry forward

Goodness of fit vs. independence · **expected counts** · χ^2 **statistic** · **ANOVA** · **F-statistic** · **between- vs. within-group variation** · **post-hoc / multiple comparisons** · **correlation** r · **least-squares line** · **slope inference** · R^2 · **residual** · **extrapolation**.

Common pitfalls

- **Chi-square needs adequate expected counts** (commonly each ≥ 5) — not observed counts.
- **A significant ANOVA doesn't say which groups differ** — that's what post-hoc comparisons are for, and running many naive t-tests inflates the Type I error.
- **Correlation captures only *linear* association** — a strong curve can have $r \approx 0$.
- R^2 **is the fraction of variation explained**, not the slope and not a probability.
- **Don't extrapolate** — a regression line is only trustworthy within the range of the observed data.
- **A real (discernible) slope is not necessarily a large or useful one** — check the size, not just the p-value.

Exercises

The Unit 4 skill is **reading the variables and choosing the right method.** Answers are provided so you can check your work.

1. **Which method?** For each scenario, name the procedure (chi-square goodness of fit, chi-square independence, ANOVA, correlation, or linear regression).

   a. Is eye color associated with whether a person is left- or right-handed?
   b. Do mean test scores differ among students taught by four different methods?
   c. Does a die's roll distribution match the expected uniform $1/6$ for each face?
   d. How strongly, and in what direction, are height and arm span related?
   e. Can we predict a house's price from its square footage, and is the relationship real?

   Show answer

   a. chi-square **test of independence** (two categorical). **b. ANOVA** (one numerical outcome, four groups). **c.** chi-square **goodness of fit** (one categorical vs. a claimed distribution). **d. correlation** (two numerical, strength/direction). **e. linear regression** (predict + test the slope).

2. **Two groups or many?** A study compares a numerical outcome (recovery time) across treatment groups. When does this call for a **two-sample t** (Unit 3) and when for **ANOVA** (Unit 4)?

   Show answer

Two groups → two-sample t. Three or more groups → ANOVA (the F-test), because running all pairwise t-tests would inflate the overall Type I error. A significant ANOVA is then followed by **multiple-comparison** procedures to locate the differences.

3. **After a significant F.** An ANOVA on five diets yields a small p-value. A student concludes "diet 1 and diet 5 differ." Is that conclusion justified by the ANOVA alone?

   Show answer

   No. A significant F only says **at least one** pair of diet means differs — not which. To claim diet 1 vs. diet 5 specifically, use a **multiple-comparison** procedure that controls the family-wise error across all the pairs.

4. **Correlation gotcha.** Two numerical variables show a strong U-shaped relationship, yet $r \approx 0$. Is "no relationship" the right read?

   Show answer

   No. r measures only **linear** association. A strong **curved** pattern can produce r near 0 — always look at the **scatterplot**, not just the correlation.

5. **Don't over-trust the line.** A regression of crop yield on rainfall is fit from seasons with 10–30 inches of rain. A planner uses it to predict yield at **60 inches**. What's the problem?

   Show answer

   Extrapolation. 60 inches is far outside the observed 10–30 inch range, where the linear pattern may not hold at all (too much rain could *reduce* yield). Predictions are only trustworthy **within** the range of the data used to fit the line.

Tech Tutorial: Apply

Tech Tutorial — Apply. This unit extends inference to more than two groups and to relationships between variables: **chi-square**, **ANOVA**, and **regression**. Pick **one** tool and follow its tab.

We reuse `class_survey.csv`.

What you'll do

1. A **chi-square test of independence** — are `year` and `housing` associated?
2. A **one-way ANOVA** — does mean `sleep_hours` differ across class `year`?
3. A **simple linear regression** — does `study_hours` predict `sleep_hours`?

---

1. Chi-square test of independence

Both variables are categorical, so we cross-tabulate `year` × `housing` and test whether the two are independent.

23.11.1 StatLens

1. Open the Chi-Square Independence tool.
2. Confirm the two variables are `year` and `housing`.
3. Read the **contingency table, expected counts, χ^2 statistic, df, and p-value**.

23.11.2 Jamovi

1. Open `class_survey.csv`.
2. **Frequencies → Independent Samples (χ^2 test of association).** Rows: `year`; Columns: `housing`.
3. Under **Statistics**, read χ^2 and **p**; under **Cells**, request **Expected counts**.

(*Jamovi screenshots to be added.*)

23.11.3 R

```
library(readr)
library(dplyr)
library(infer)

survey <- read_csv("../datasets/class_survey.csv")
```

```
survey |> chisq_test(housing ~ year)
```

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl>    <int>   <dbl>
1      4.62        3   0.202
```

> ⚠ Check the conditions
>
> With only 50 students split across a 4×2 table, some **expected counts fall below 5** — the chi-square approximation is shaky here. In practice you'd report that caveat (and could fall back to the *randomization* chi-square from the Simulate unit). For this tutorial the point is the mechanics.

---

2. One-way ANOVA

ANOVA compares the mean of a numerical variable (`sleep_hours`) across 3+ groups (the four class years) with a single *F*-test.

23.11.4 StatLens

1. Open the One-Way ANOVA tool.
2. Confirm **response** `sleep_hours`, **grouping** `year`.
3. Read the group means, the **F statistic**, df, and **p-value**.

23.11.5 Jamovi

1. **ANOVA → One-Way ANOVA.** Dependent: `sleep_hours`; Grouping: `year`.
2. Read the **F**, **df**, and **p**. (Use **Descriptives plots** to see the group means.)

(*Jamovi screenshots to be added.*)

23.11.6 R

```
aov(sleep_hours ~ year, data = survey) |> summary()
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
year         3   2.78  0.9266   1.119  0.351
Residuals   46  38.08  0.8278
```

---

3. Simple linear regression

Both variables are numerical. We fit a line predicting `sleep_hours` from `study_hours` and test whether the slope differs from zero. (In the Explore unit the scatter of these two looked flat — regression now tests that formally.)

23.11.7 StatLens

1. Open the Regression Slope tool.
2. Set **x** to `study_hours` and **y** to `sleep_hours`.
3. Read the fitted **slope**, its **p-value**, and **R^2**.

23.11.8 Jamovi

1. **Regression → Linear Regression.** Dependent: `sleep_hours`; Covariate: `study_hours`.
2. Read the **slope** (the `study_hours` coefficient), its **p**, and the **R^2** (Model Fit).

(*Jamovi screenshots to be added.*)

23.11.9 R

```
lm(sleep_hours ~ study_hours, data = survey) |> summary()
```

```
Call:
lm(formula = sleep_hours ~ study_hours, data = survey)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1533 -0.5092 -0.1111  0.5444  1.9448

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.14892    0.30485  23.450   <2e-16 ***
study_hours -0.00625    0.01982  -0.315    0.754
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9217 on 48 degrees of freedom
Multiple R-squared:  0.002068,  Adjusted R-squared:  -0.01872
F-statistic: 0.09947 on 1 and 48 DF,  p-value: 0.7538
```

Check yourself

- ☐ run a chi-square test of independence and check its expected counts;
- ☐ run a one-way ANOVA and report F, df, and p;
- ☐ fit a simple regression and report the slope, its p-value, and R^2;
- ☐ match each procedure to the data shape that calls for it.

Going further in R (optional)

OpenIntro *IMS* interactive tutorials for this unit:

- Chi-squared test of independence
- Comparing many means (ANOVA)
- Simple linear regression
- Inference for regression (t-test for slope)

Project: Comparing Several Means (ANOVA)

Project — Analysis of Variance. Carry out the analysis in **Jamovi** (or **R**) and **communicate** the result. ANOVA asks whether *any* of several group means differ; the graded skills are framing that question, checking conditions, and interpreting the F-test in context.

The question

In baseball, a player's **on-base percentage** (OBP) measures how often they reach base. Do players at different field positions tend to get on base at different rates?

> **Do mean OBP values differ across infielders, outfielders, and catchers?**

The data

`mlb_players_18` holds 2018 Major League Baseball players. You'll use `OBP` and `position`. We restrict to **position players with regular playing time** (at least 100 at-bats) and group positions into **Infield**, **Outfield**, and **Catcher** (pitchers and designated hitters are set aside). The R tab builds that grouping; in Jamovi you would create the same grouped column.

Plan it first

a. **Variables.** Identify the numerical **response** and the categorical **explanatory** variable, and how many groups it has.
b. **Hypotheses.** State H_0 (all group means equal) and H_A (at least one differs). Why is H_A *not* "all means are different"?
c. **Procedure.** One numerical outcome across **three** groups → **one-way ANOVA** (the F-test), not a pile of t-tests. Why not just run all the pairwise t-tests?
d. **After ANOVA.** If F is discernible, what additional step tells you *which* positions differ?

Do the analysis

23.11.10 Jamovi

1. Open `mlb_players_18.csv`; filter to `AB >= 100` and create a `pos_group` column (Infield / Outfield / Catcher).
2. **ANOVA → One-Way ANOVA.** Dependent variable: `OBP`; Grouping variable: `pos_group`.
3. Request the **F-test**, **descriptives**, a **homogeneity (Levene)** check, and **post-hoc** comparisons (Tukey).
4. Read off the group means, the F-statistic and p-value, and which pairs the post-hoc test flags.

(*Jamovi screenshots to be added.*)

23.11.11 R

```
library(infer)
library(dplyr)
library(ggplot2)
library(openintro)
data(mlb_players_18)

# Regular position players, grouped into Infield / Outfield / Catcher
infield  <- c("1B", "2B", "3B", "SS")
outfield <- c("LF", "CF", "RF")
d <- mlb_players_18 |>
  filter(position %in% c(infield, outfield, "C"), AB >= 100) |>
  mutate(pos_group = case_when(
    position %in% infield  ~ "Infield",
    position %in% outfield ~ "Outfield",
    TRUE                   ~ "Catcher"))

# Group means and sizes
d |>
  group_by(pos_group) |>
  summarize(mean_OBP = mean(OBP), sd = sd(OBP), n = n())
```

```
# A tibble: 3 x 4
  pos_group mean_OBP     sd     n
  <chr>        <dbl>  <dbl> <int>
1 Catcher      0.302 0.0382    64
2 Infield      0.318 0.0379   205
3 Outfield     0.320 0.0426   160
```

```
# Observed F statistic, then a randomization (permutation) F distribution
obs_F <- d |> specify(OBP ~ pos_group) |> calculate(stat = "F")
obs_F
```

```
Response: OBP (numeric)
Explanatory: pos_group (factor)
# A tibble: 1 x 1
   stat
  <dbl>
1  5.08
```

```
set.seed(145)
null_F <- d |>
  specify(OBP ~ pos_group) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "F")

null_F |> get_p_value(obs_stat = obs_F, direction = "greater")
```

```
# A tibble: 1 x 1
```

```
  p_value
    <dbl>
1   0.006
```

```
ggplot(d, aes(pos_group, OBP, fill = pos_group)) +
  geom_boxplot(show.legend = FALSE) +
  labs(x = "Position group", y = "On-base percentage") +
  theme_minimal(base_size = 13)
```

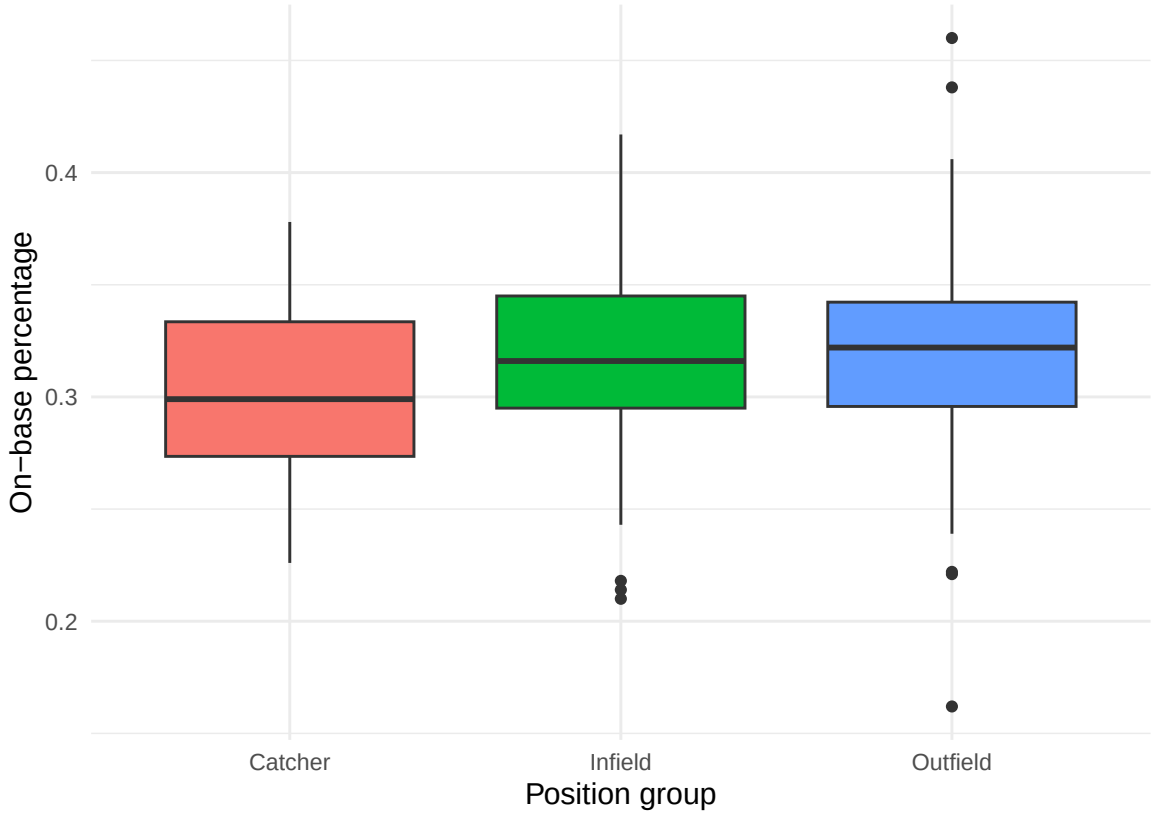


Figure 23.1: On-base percentage by position group for regular MLB position players in 2018.

Your deliverable

Submit a short report containing **all** of:

a. **Graphic** — labeled boxplots of `OBP` by position group.
b. **Numerical result** — the group means, the **F-statistic**, and the p-value.
c. **Conditions** — one sentence on the ANOVA conditions (independent groups, roughly normal, similar spreads).
d. **Conclusion in context** — 3–4 sentences: is there **discernible** evidence that mean OBP differs by position group? If so, note that ANOVA alone doesn't say *which* groups differ, and report what the post-hoc comparison found (if you ran it).
e. **One limitation** — e.g., position grouping is a simplification; 2018 only.

How it's graded

Criterion	What we look for
Appropriate, labeled graphic	boxplots by group, axes labeled
Procedure & conditions	one-way ANOVA identified; conditions addressed
Correct numerical result	group means, F, p-value reported correctly
Conclusion **in context**	discernibility + "F doesn't say which pair" + post-hoc if used
Communication	clear, correct interpretation

Warm up first. Build intuition for between- vs. within-group variation in the ANOVA tool before writing your report.

Project: Association Between Two Categorical Variables

Project — Chi-Square Test of Independence. Carry out the analysis in **Jamovi** (or **R**) and **communicate** the result. The graded skills are framing the association question, checking the expected-count condition, and interpreting the test in context.

The question

Few airplane debates are as heated as whether it's acceptable to **recline your seat**. FiveThirtyEight ran a national survey on flying etiquette, asking travelers whether reclining is rude — and recording each respondent's age.

> **Is a flyer's opinion on reclining associated with their age?**

The data

`flying_etiquette.csv` holds the survey responses (about 1,040 people; ~843 answered both questions here). You'll use two categorical variables: `recline_rude` (`No`, `Somewhat`, `Very` — how rude is it to recline?) and `age` (`18-29`, `30-44`, `45-60`, `> 60`). Source: FiveThirtyEight, *"Hey, You Should Stop Reclining Your Airplane Seat."*

Plan it first

a. **Variables.** Confirm both variables are categorical and give the size of the table (rows × columns).
b. **Hypotheses.** State H_0 and H_A for a test of **independence** in words (opinion on reclining and age are independent vs. associated).
c. **Eyeball it.** Look at the row percentages (the opinion mix within each age group). Can you tell *for sure* whether opinion depends on age, or could the differences be chance? This is the judgment the test makes rigorous.
d. **Procedure & condition.** Two categorical variables, "are they associated?" → **chi-square test of independence**. It needs **expected counts** that aren't too small (a common rule: each ≥ 5) — what will you check?

Do the analysis

23.11.12 Jamovi

1. Open `flying_etiquette.csv`.
2. **Frequencies → Independent Samples (χ² test of association).** Rows: `age`; Columns: `recline_rude`.
3. Request the **χ² test**, **expected counts**, and **row percentages**.
4. Read off the table, confirm expected counts are large enough, and record χ^2 and the p-value.

(*Jamovi screenshots to be added.*)

23.11.13 R

```
library(infer)
library(dplyr)
library(readr)
library(ggplot2)

flying <- read_csv("../datasets/flying_etiquette.csv")

# Keep respondents who answered both questions; order the categories
fly <- flying |>
  filter(!is.na(recline_rude), !is.na(age)) |>
  mutate(age          = factor(age, levels = c("18-29", "30-44", "45-60", "> 60")),
         recline_rude = factor(recline_rude, levels = c("No", "Somewhat", "Very")))

# Contingency table
tab <- table(fly$age, fly$recline_rude)
tab
```

```
         No Somewhat Very
  18-29  78       74   20
  30-44 143       64   15
  45-60 140       80   14
  > 60  133       61   21
```

```
# Row percentages: the opinion mix within each age group (try to eyeball it)
round(100 * prop.table(tab, 1), 0)
```

```
       No Somewhat Very
  18-29 45       43   12
  30-44 64       29    7
  45-60 60       34    6
  > 60  62       28   10
```

```
# Chi-square test of independence (infer)
fly |> chisq_test(recline_rude ~ age)
```

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl>    <int>   <dbl>
1      19.8        6 0.00305
```

```
# Condition check: smallest expected count (rule of thumb: each >= 5)
cat("Smallest expected count:", round(min(chisq.test(tab)$expected), 1), "\n")
```

```
Smallest expected count: 14.3
```

```
ggplot(fly, aes(age, fill = recline_rude)) +
  geom_bar(position = "fill") +
  labs(x = "Age group", y = "Proportion", fill = "Reclining is rude?") +
  theme_minimal(base_size = 13)
```

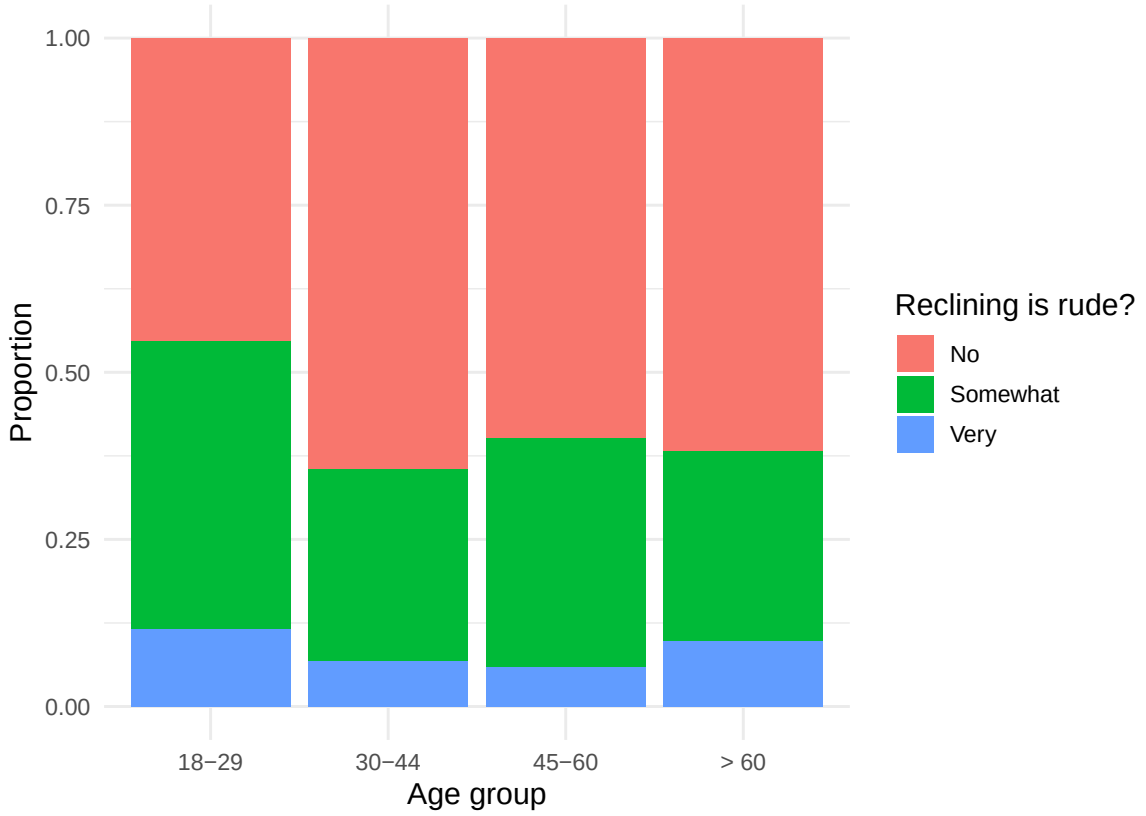


Figure 23.2: Opinion on reclining by age group. Younger flyers are the most likely to call reclining rude — but is the difference real or noise?

Your deliverable

Submit a short report containing **all** of:

a. **Graphic** — a labeled stacked (or side-by-side) bar chart showing the opinion mix by age group.
b. **Numerical result** — the contingency table, the χ^2 **statistic**, the degrees of freedom, and the p-value.
c. **Conditions** — one sentence confirming the expected counts are large enough for the chi-square approximation.
d. **Conclusion in context** — 3–4 sentences: is there **discernible** evidence that opinion on reclining is associated with age? Describe the pattern (who finds it rudest?), note how the test settled what the percentages alone left uncertain, and stress that association is **not** causation.
e. **One limitation** — e.g., this is a single online survey; opinions may shift over time or differ by who chose to respond.

How it's graded

Criterion	What we look for
Appropriate, labeled graphic	opinion mix by group, axes/legend labeled
Procedure & conditions	chi-square independence identified; expected counts checked
Correct numerical result	table, χ^2, df, p-value reported correctly
Conclusion **in context**	discernibility + the age pattern + association ≠ causation
Communication	clear, correct interpretation

Warm up first. Explore expected counts and the χ^2 statistic in the Chi-Square tool before writing your report.

Project: Association Between Two Categorical Variables (AI version)

Project — Chi-Square Test of Independence. Carry out the analysis in **Jamovi** (or **R**) and **communicate** the result. The graded skills are framing the association question, checking the expected-count condition, and interpreting the test in context.

The question

AI coding tools (ChatGPT, Copilot, and the like) swept into software work almost overnight. A natural question: is adoption uniform, or does it depend on who you are?

> **Among developers, is using AI tools associated with age?**

The data

`ai_tool_use.csv` is a **random sample of 600 developers** from the Stack Overflow Annual Developer Survey (2024). You'll use two categorical variables: `ai_use` (`Yes`, `Not yet, but plan to`, `No` — do you use AI tools in your development work?) and `age_group` (`18-24`, `25-34`, `35-44`, `45+`).

> **Data source & license.** Stack Overflow Annual Developer Survey 2024 (via the TidyTuesday project), used under the **Open Database License (ODbL) 1.0**. This excerpt is a derived subset; it is shared under ODbL — an exception to this book's CC BY-SA license.

Plan it first

a. **Variables.** Confirm both variables are categorical and give the size of the table (rows × columns).
b. **Hypotheses.** State H_0 and H_A for a test of **independence** in words (AI-tool use and age are independent vs. associated).
c. **Eyeball it.** Look at the row percentages (the AI-use mix within each age group). Can you tell *for sure* whether adoption depends on age, or could the differences be chance? This is the judgment the test makes rigorous.
d. **Procedure & condition.** Two categorical variables, "are they associated?" → **chi-square test of independence**. It needs **expected counts** that aren't too small (a common rule: each ≥ 5) — what will you check?

Do the analysis

23.11.14 Jamovi

1. Open `ai_tool_use.csv`.
2. **Frequencies → Independent Samples (χ² test of association).** Rows: `age_group`; Columns: `ai_use`.
3. Request the **χ² test**, **expected counts**, and **row percentages**.
4. Read off the table, confirm expected counts are large enough, and record χ^2 and the p-value.

(*Jamovi screenshots to be added.*)

23.11.15 R

```
library(infer)
library(dplyr)
library(readr)
library(ggplot2)

ai <- read_csv("../datasets/ai_tool_use.csv")

# Order the categories
ai <- ai |>
  mutate(age_group = factor(age_group, levels = c("18-24", "25-34", "35-44", "45+")),
         ai_use     = factor(ai_use, levels = c("Yes", "Not yet, but plan to", "No")))

# Contingency table
tab <- table(ai$age_group, ai$ai_use)
tab
```

```
       Yes Not yet, but plan to  No
  18-24  96                   16  23
  25-34 136                   19  44
  35-44  94                   26  42
  45+    51                   24  29
```

```
# Row percentages: the AI-use mix within each age group (try to eyeball it)
round(100 * prop.table(tab, 1), 0)
```

```
       Yes Not yet, but plan to No
  18-24  71                   12 17
  25-34  68                   10 22
  35-44  58                   16 26
  45+    49                   23 28
```

```
# Chi-square test of independence (infer)
ai |> chisq_test(ai_use ~ age_group)
```

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl>    <int>   <dbl>
1      19.8        6 0.00306
```

```
# Condition check: smallest expected count (rule of thumb: each >= 5)
cat("Smallest expected count:", round(min(chisq.test(tab)$expected), 1), "\n")
```

```
Smallest expected count: 14.7
```

```
ggplot(ai, aes(age_group, fill = ai_use)) +
  geom_bar(position = "fill") +
  labs(x = "Age group", y = "Proportion", fill = "Uses AI tools?") +
  theme_minimal(base_size = 13)
```

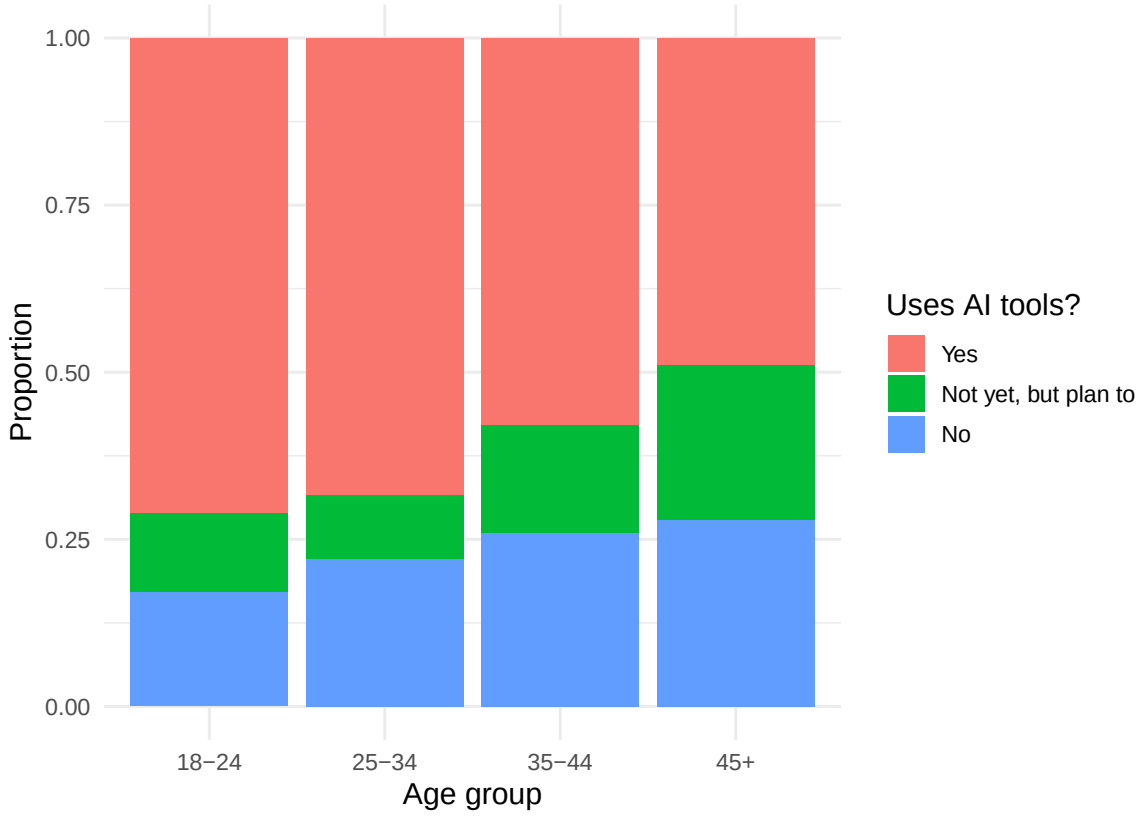


Figure 23.3: AI-tool use by age group in a sample of 600 developers. Younger developers adopt AI tools more — but is the difference real or noise?

Your deliverable

Submit a short report containing **all** of:

a. **Graphic** — a labeled stacked (or side-by-side) bar chart showing the AI-use mix by age group.
b. **Numerical result** — the contingency table, the χ^2 **statistic**, the degrees of freedom, and the p-value.
c. **Conditions** — one sentence confirming the expected counts are large enough for the chi-square approximation.
d. **Conclusion in context** — 3–4 sentences: is there **discernible** evidence that AI-tool use is associated with age? Describe the pattern (who adopts most?), note how the test settled what the percentages alone left uncertain, and stress that association is **not** causation.

e. **One limitation** — e.g., this samples developers who chose to take one survey; results may not generalize to all workers, and attitudes change fast.

How it's graded

Criterion	What we look for
Appropriate, labeled graphic	AI-use mix by group, axes/legend labeled
Procedure & conditions	chi-square independence identified; expected counts checked
Correct numerical result	table, χ^2, df, p-value reported correctly
Conclusion **in context**	discernibility + the age pattern + association ≠ causation
Communication	clear, correct interpretation

Warm up first. Explore expected counts and the χ^2 statistic in the Chi-Square tool before writing your report.

Project: Fitting and Testing a Regression Line

Project — Linear Regression. Carry out the analysis in **Jamovi** (or **R**) and **communicate** the result. The graded skills are framing the relationship, fitting and *interpreting* the line, testing whether the slope is real, and not over-trusting the model.

The question

Field biologists often need to estimate an animal's overall size from a part that's easy to measure. For brushtail possums, the tail is simple to measure in the field.

> **Can we predict a possum's total body length from its tail length, and is the relationship real?**

The data

The `possum` dataset records 104 brushtail possums. You'll use `tail_l` (tail length, cm) as the explanatory variable and `total_l` (total body length, cm) as the response.

Plan it first

a. **Variables.** Identify the explanatory and response variables (both numerical).
b. **Display.** What plot shows the relationship between two numerical variables?
c. **Procedure.** To predict one numerical variable from another and test whether the association is real → **linear regression** with a **t-test on the slope**.
d. **Hypotheses.** State H_0 and H_A for the slope. What does the slope being 0 mean?

Do the analysis

23.11.16 Jamovi

1. Open `possum.csv`.
2. **Regression → Linear Regression.** Dependent variable: `total_l`; Covariate: `tail_l`.
3. Request the **model coefficients** (estimate, t, p), R^2, and a **scatterplot with the fitted line**; check the **residual** plots.
4. Read off the slope, its p-value, and R^2.

(*Jamovi screenshots to be added.*)

23.11.17 R

```
library(infer)
library(dplyr)
library(ggplot2)
library(openintro)
data(possum)

# Fit the least-squares line (intercept and slope)
possum |> specify(total_l ~ tail_l) |> fit()
```

```
# A tibble: 2 x 2
  term      estimate
  <chr>        <dbl>
1 intercept    41.0
2 tail_l        1.24
```

```
# Correlation and R^2 (for simple regression, R^2 = r^2)
r <- cor(possum$tail_l, possum$total_l)
cat("r =", round(r, 3), " R^2 =", round(r^2, 3), "\n")
```

```
r = 0.566  R^2 = 0.32
```

```
# Permutation test for the slope: is it discernibly different from 0?
obs_fit <- possum |> specify(total_l ~ tail_l) |> fit()

set.seed(145)
null_fits <- possum |>
  specify(total_l ~ tail_l) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  fit()

null_fits |>
  get_p_value(obs_stat = obs_fit, direction = "two-sided") |>
  filter(term == "tail_l")
```

```
# A tibble: 1 x 2
  term   p_value
  <chr>    <dbl>
1 tail_l       0
```

```
ggplot(possum, aes(tail_l, total_l)) +
  geom_point(alpha = 0.6, color = "#3E5496") +
  geom_smooth(method = "lm", se = TRUE, color = "#C7254E") +
  labs(x = "Tail length (cm)", y = "Total length (cm)") +
  theme_minimal(base_size = 13)
```

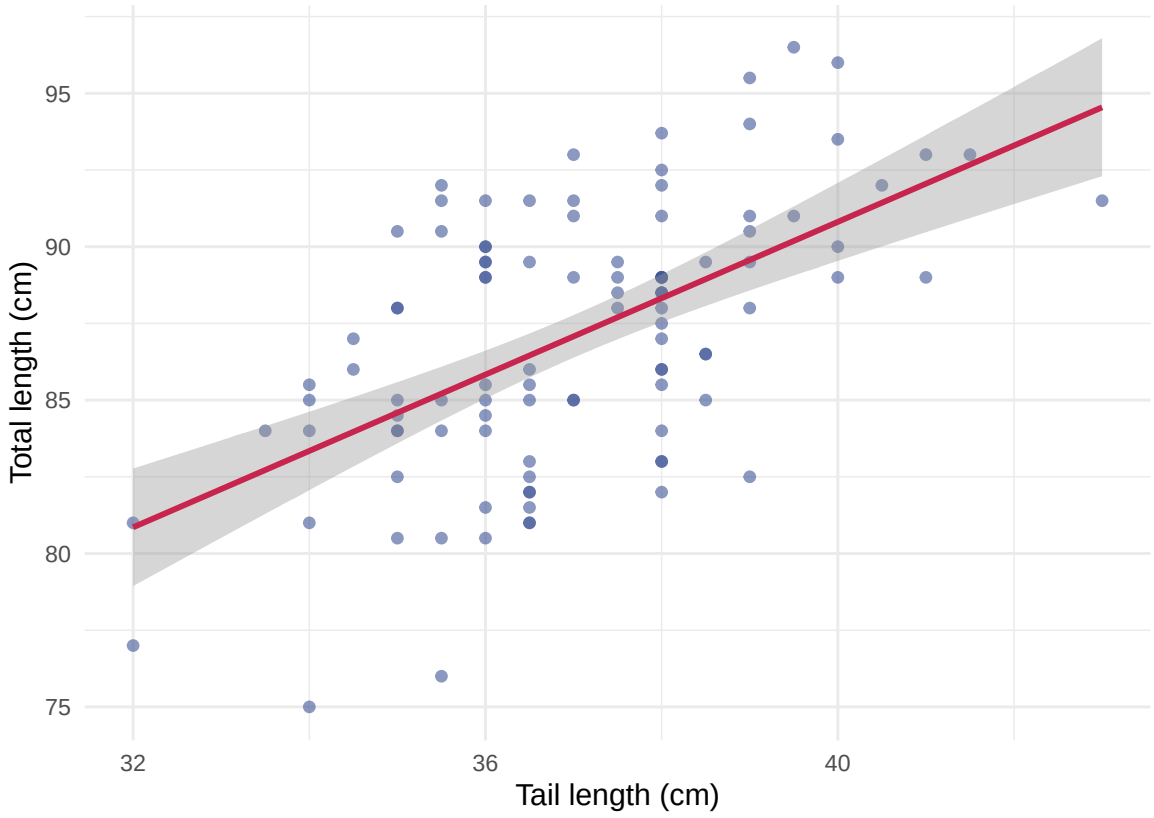


Figure 23.4: Total body length vs. tail length for 104 brushtail possums, with the least-squares line.

> **Note**
>
> **Reading the permutation p-value.** The relationship is strong enough that *none* of the shuffled slopes reaches the observed slope, so the simulation returns a p-value of **0**. Report that honestly as **p < 0.001** (smaller than $1/\text{reps}$) — very strong evidence — not literally zero.

Your deliverable

Submit a short report containing **all** of:

a. **Graphic** — a labeled scatterplot of `total_l` vs. `tail_l` with the fitted line.
b. **Numerical result** — the regression equation (intercept and slope), the slope's p-value, and R^2.
c. **Interpret the slope** — one sentence in context: how much does predicted total length change per extra cm of tail?
d. **Conclusion in context** — 3–4 sentences: is the slope **discernibly** different from 0? How strong is the relationship (R^2)? Would you trust a prediction for a tail length far outside the observed range (**extrapolation**)?
e. **One limitation** — e.g., one species/region; correlation here is descriptive, not causal.

How it's graded

Criterion	What we look for
Appropriate, labeled graphic	scatterplot with fitted line, axes labeled
Procedure & conditions	regression + slope test identified; residuals considered
Correct numerical result	equation, slope p-value, R^2 reported correctly
Conclusion **in context**	slope interpreted; discernibility; no extrapolation
Communication	clear, correct interpretation

Warm up first. Fit a line by eye and see the residuals in the Regression explorer before writing your report.

PART V

UNIT 5: FOUNDATIONS

You have already been *doing* probability — every simulation you ran was an experiment in randomness. This part formalizes the rules governing random events, introduces random variables, and develops the binomial and normal distributions as mathematical models. These foundations explain *why* the methods from earlier parts work.

The final chapters explore what happens when inference goes wrong: Type II errors, statistical power, and how to plan studies large enough to detect meaningful effects. These topics are essential for designing good research and critically evaluating published results.

Chapter 24

Probability Rules

Probability forms the foundation of statistics, and you are probably already aware of many of the ideas we will discuss in this chapter. However, formalizing these concepts is likely new. While this chapter provides the theoretical foundation for the inference methods you have already been using, it also offers practical tools for reasoning about uncertainty. We will define probability, develop rules for combining probabilities, and explore conditional probability and independence — all concepts that underpin the inference methods from Parts II and III.

24.1 Defining probability

Statistics is built on probability, and probability gives us the language to describe uncertainty. We begin with a few simple examples that may feel familiar.

A "die," the singular of "dice," is a cube with six faces numbered 1, 2, 3, 4, 5, and 6. What is the chance of getting a 1 when rolling a fair die?

If the die is fair, then the chance of a 1 is as good as the chance of any other number. Since there are six equally likely outcomes, the chance must be 1-in-6 or, equivalently, $1/6$.

What is the chance of getting a 1 or 2 on the next roll?

The outcomes 1 and 2 constitute two of the six equally likely outcomes, so the chance of getting one of these two is $2/6 = 1/3$.

What is the chance of not rolling a 2?

Since the chance of rolling a 2 is $1/6$, the chance of not rolling a 2 must be $1 - 1/6 = 5/6$. Alternatively, not rolling a 2 means getting a 1, 3, 4, 5, or 6, which makes up five of the six equally likely outcomes.

We use probability to describe and understand apparent randomness. We frame probability in terms of a **random process** giving rise to an **outcome**.

Random process	Possible outcomes
Roll a die	1, 2, 3, 4, 5, or 6
Flip a coin	H or T

The **probability** of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times. Probability always takes values between 0 and 1 (inclusive) and may also be expressed as a percentage between 0% and 100%.

Probability can be illustrated by rolling a die many times. Let $\hat{p}_n$ be the proportion of outcomes that are 1 after the first n rolls. As the number of rolls increases, $\hat{p}_n$ converges to the probability $p = 1/6$.

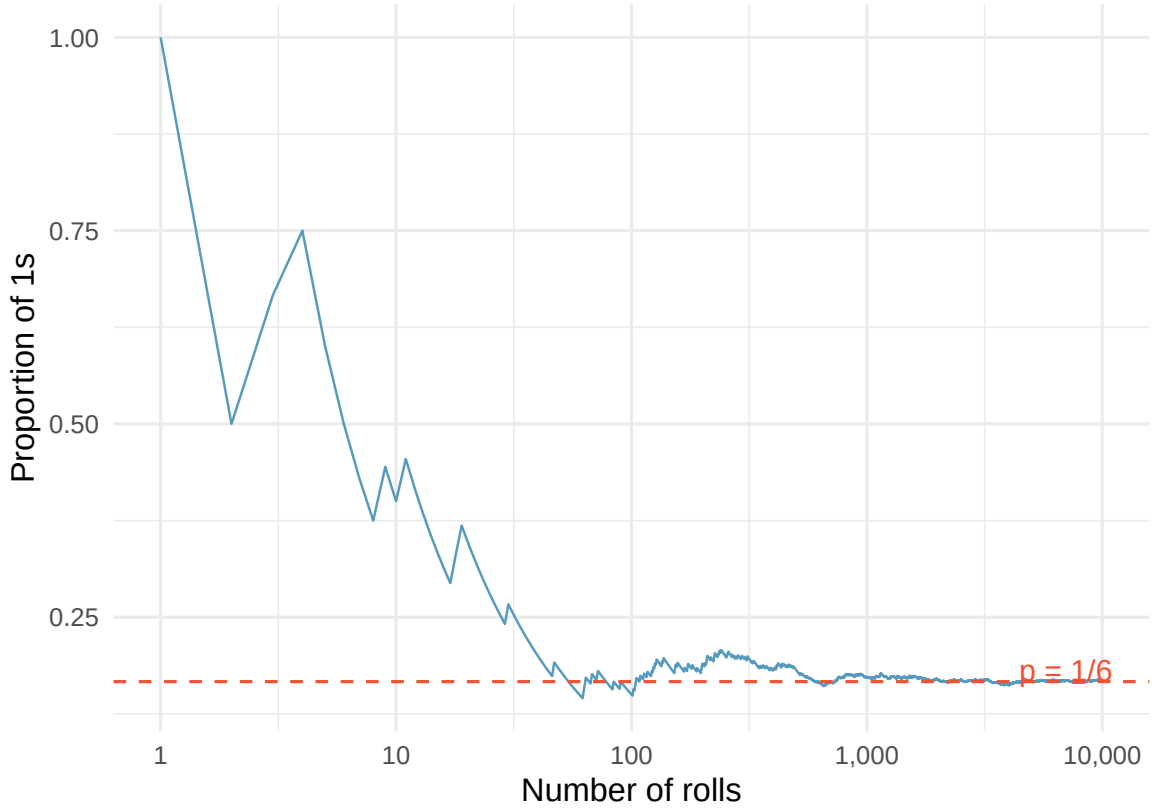


Figure 24.1: The fraction of die rolls that are 1 at each stage in a simulation. The proportion tends to get closer to 1/6 as the number of rolls increases.

Law of Large Numbers. As more observations are collected, the proportion $\hat{p}_n$ of occurrences with a particular outcome converges to the probability p of that outcome.

Occasionally the proportion will veer away from the true probability, but these deviations become smaller as the number of trials increases.

We write the probability of rolling a 1 as $P(\text{rolling a } 1)$. When the context is clear, we abbreviate this as $P(1)$.

Random processes include rolling a die and flipping a coin. (a) Think of another random process. (b) Describe all the possible outcomes of that process.

Show answer

(a) and (b) Here are several examples. (i) Whether someone gets sick next month: outcomes are *sick* and *not sick*. (ii) Randomly selecting a person and measuring their height: the outcome is a positive number. (iii) Whether the stock market goes up or down next week: outcomes are *up*, *down*, and *no change*. (iv) The number of emails you receive tomorrow: the outcome is a non-negative integer.

24.1.1 Simulation: The law of large numbers in action

One of the most powerful ways to build intuition about probability is through simulation. Imagine writing a computer program that simulates flipping a fair coin 10,000 times and tracking the running proportion of heads after each flip.

- After 10 flips, the proportion of heads might be 0.30 or 0.70 — quite far from 0.50.
- After 100 flips, the proportion is likely between 0.40 and 0.60.
- After 1,000 flips, the proportion is usually between 0.47 and 0.53.
- After 10,000 flips, it is within 0.01 of 0.50 about 95% of the time.

This is the Law of Large Numbers at work. The key insight is that probability is a *long-run* concept. In the short run, anything can happen. In the long run, patterns emerge with remarkable stability.

This simulation perspective will recur throughout our course. In Parts II and III, we used simulation to build sampling distributions and test hypotheses. Here in Unit 5, we develop the *mathematical* rules that explain *why* those simulations work.

24.2 Sample spaces and events

The **sample space** S of a random process is the set of all possible outcomes. An **event** is a subset of the sample space — that is, a collection of one or more outcomes.

For rolling a single die, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$.

Let A represent the event that the die roll results in 1 or 2, so $A = \{1, 2\}$. Let B represent the event that the die roll is 4 or 6, so $B = \{4, 6\}$. These events are shown visually in the diagram below.

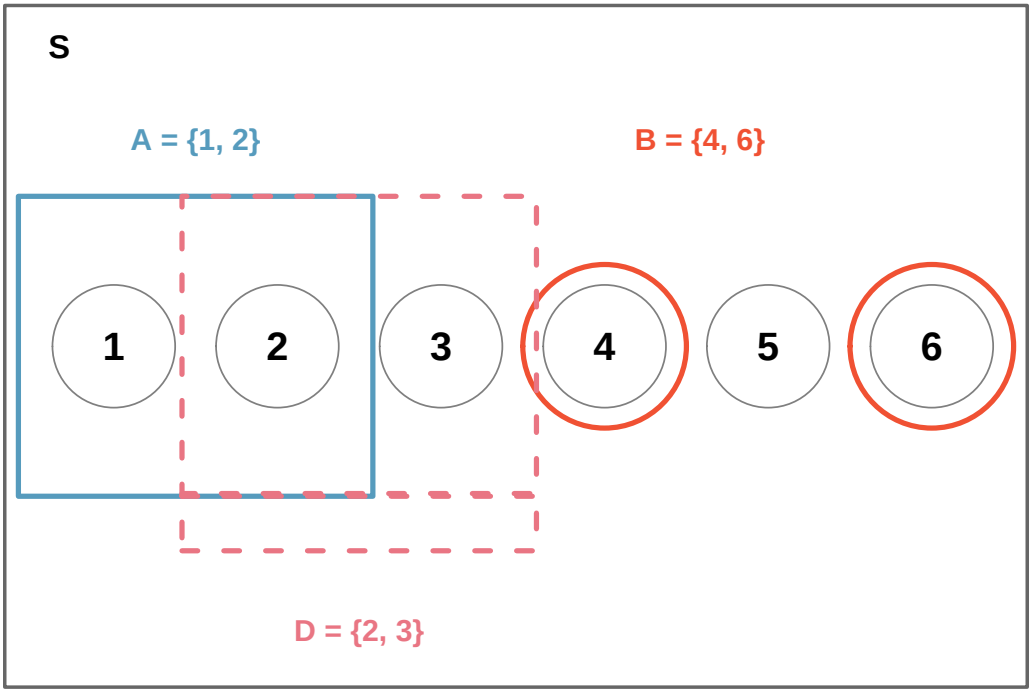


Figure 24.2: Three events for die outcomes. A and B are disjoint since they share no outcomes.

24.3 The complement rule

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is 2 or 3. Then the **complement** of D, denoted D^c, represents all outcomes in the sample space that are *not* in D:

$$D^c = \{1, 4, 5, 6\}$$

(a) Compute $P(D^c) = P(\text{rolling a 1, 4, 5, or 6})$. (b) What is $P(D) + P(D^c)$?

Show answer

(a) The outcomes are disjoint and each has probability $1/6$, so $P(D^c) = 4/6 = 2/3$. (b) $P(D) = 1/6 + 1/6 = 1/3$. Since D and D^c are disjoint, $P(D) + P(D^c) = 1/3 + 2/3 = 1$.

A complement A^c is constructed so that (i) every possible outcome not in A is in A^c, and (ii) A and A^c are disjoint. This gives us:

Complement rule. The complement of event A is denoted A^c and represents all outcomes not in A. The probabilities of A and A^c are related by:

$$P(A) + P(A^c) = 1 \qquad \text{equivalently,} \qquad P(A) = 1 - P(A^c)$$

The complement rule is especially useful when computing a probability directly would be tedious but the complement is simple.

Let A represent the event where the sum of two dice is less than 12. What is $P(A)$?

The complement A^c is the event that the sum equals 12. The only way to roll a sum of 12 is to get two 6s, so $P(A^c) = 1/36$. Using the complement rule:

$$P(A) = 1 - P(A^c) = 1 - \frac{1}{36} = \frac{35}{36}$$

Find the following probabilities for rolling two dice: (a) The sum is *not* 6. (b) The sum is at least 4. (c) The sum is no more than 10.

Show answer

(a) $P(\text{sum} = 6) = 5/36$, so $P(\text{not } 6) = 1 - 5/36 = 31/36$. (b) $P(\text{sum} = 2 \text{ or } 3) = 1/36 + 2/36 = 1/12$, so $P(\text{at least } 4) = 1 - 1/12 = 11/12$. (c) $P(\text{sum} = 11 \text{ or } 12) = 2/36 + 1/36 = 1/12$, so $P(\text{no more than } 10) = 1 - 1/12 = 11/12$.

24.4 The addition rule

24.4.1 Disjoint (mutually exclusive) events

Two outcomes or events are called **disjoint** (or **mutually exclusive**) if they cannot both happen at the same time. For instance, when rolling a die, the outcomes 1 and 2 are disjoint since they cannot both occur on a single roll.

Addition rule for disjoint events. If A_1 and A_2 are disjoint events, then:

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2)$$

More generally, if $A_1, A_2, \ldots, A_k$ are all mutually disjoint, then:

$$P(A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_k) = P(A_1) + P(A_2) + \cdots + P(A_k)$$

We are interested in the probability of rolling a 1, 4, or 5. (a) Explain why the outcomes 1, 4, and 5 are disjoint. (b) Apply the addition rule to determine $P(1 \text{ or } 4 \text{ or } 5)$.

Show answer

(a) At most one of these outcomes can occur on a single die roll, so they are disjoint. (b) $P(1 \text{ or } 4 \text{ or } 5) = P(1) + P(4) + P(5) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2$.

24.4.2 Probability distributions

A **probability distribution** is a table of all disjoint outcomes and their associated probabilities. For example, the distribution for the sum of two dice:

Dice sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Rules for probability distributions. A valid probability distribution must satisfy three rules:

1. The outcomes listed must be disjoint.
2. Each probability must be between 0 and 1.
3. The probabilities must total 1.

24.4.3 The general addition rule

When events are *not* disjoint, we cannot simply add their probabilities — doing so would double-count the outcomes they share.

Consider a standard deck of 52 cards. Let A be the event "the card is a diamond" and B be the event "the card is a face card." These events overlap: the jack, queen, and king of diamonds are both diamonds and face cards.

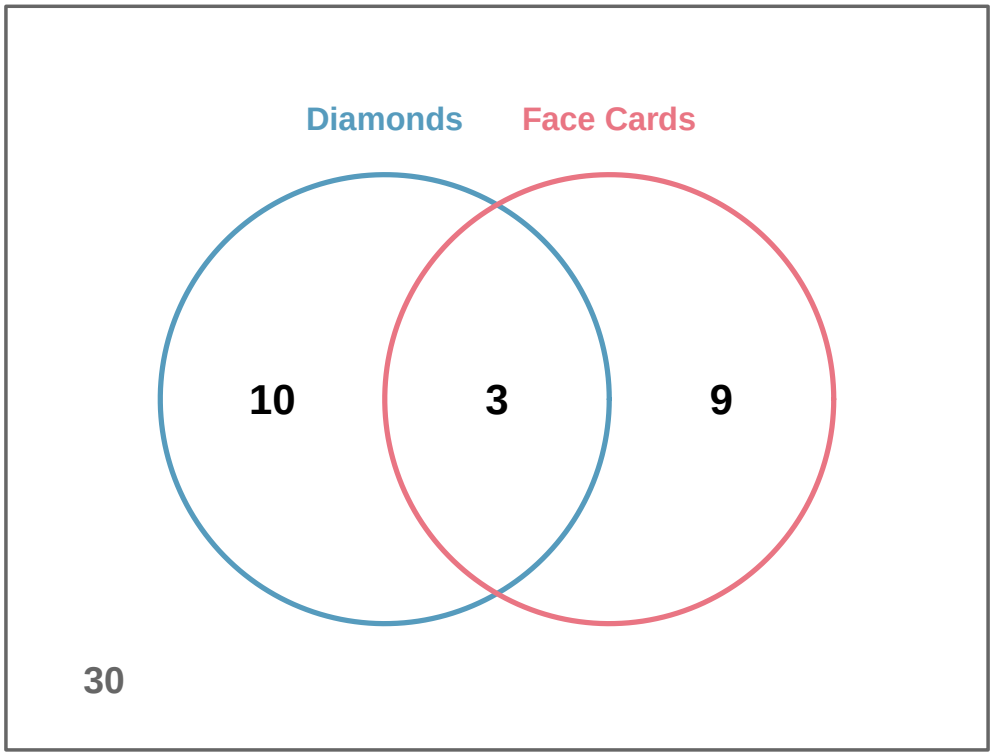


Figure 24.3: A Venn diagram for diamonds and face cards showing 10 diamonds that are not face cards, 9 face cards that are not diamonds, and 3 cards that are both.

If we compute $P(A) + P(B) = 13/52 + 12/52 = 25/52$, we have counted the 3 cards in the overlap twice. We correct this:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52} = \frac{11}{26}$$

General addition rule. For any two events A and B (disjoint or not):

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

"Or" is inclusive in statistics. When we write "A or B," we mean "A, B, or both" unless explicitly stated otherwise.

(a) If A and B are disjoint, explain why $P(A \text{ and } B) = 0$. (b) Show that the general addition rule simplifies to the simpler addition rule when A and B are disjoint.

Show answer

(a) If A and B are disjoint, they can never occur simultaneously, so $P(A \text{ and } B) = 0$. (b) Substituting 0 for $P(A \text{ and } B)$ in the general addition rule gives $P(A \text{ or } B) = P(A) + P(B) - 0 = P(A) + P(B)$, which is the simpler addition rule.

24.5 Conditional probability

24.5.1 Exploring conditional probability with data

There can be rich relationships between two or more variables, and conditional probability provides the framework for exploring them. Suppose we have data from a study classifying 1,822 photos using both a machine learning (ML) algorithm and human judgment:

	Truth: fashion	Truth: not	Total
ML: pred_fashion	197	22	219
ML: pred_not	112	1491	1603
Total	309	1513	1822

If a photo is actually about fashion, what is the chance the ML classifier correctly identified it?

Of the 309 fashion photos, the ML algorithm correctly classified 197:

$$P(\text{ML predicts fashion} \mid \text{truth is fashion}) = \frac{197}{309} = 0.638$$

We call this a **conditional probability** because we computed the probability *under a condition* — that the photo was actually about fashion.

24.5.2 Marginal and joint probabilities

A **marginal probability** is a probability based on a single variable, without regard to other variables. A **joint probability** is the probability of two or more outcomes occurring together.

From the photo classification data: $P(\text{ML predicts fashion}) = 219/1822 = 0.120$ is a marginal probability, while $P(\text{ML predicts fashion and truth is fashion}) = 197/1822 = 0.108$ is a joint probability.

24.5.3 The conditional probability formula

Conditional probability. The conditional probability of outcome A given condition B is:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

The vertical bar "|" is read as "given."

Using the photo classification data, we can verify our earlier calculation using proportions alone:

$$P(\text{truth is fashion} \mid \text{ML predicts fashion}) = \frac{P(\text{truth is fashion and ML predicts fashion})}{P(\text{ML predicts fashion})} = \frac{0.108}{0.120} = 0.900$$

If the ML classifier says a photo is about fashion, there is a 90% probability it actually is.

The smallpox dataset provides data on 6,224 individuals in Boston in 1721. Of these, 3.92% were inoculated. Among those not inoculated, 14.11% died. Among those inoculated, only 2.46% died. Write out, in formal notation, the probability a randomly selected person who was not inoculated died from smallpox.

Show answer

$P(\text{result} = \text{died} \mid \text{inoculated} = \text{no}) = \frac{P(\text{died and not inoculated})}{P(\text{not inoculated})} = \frac{0.1356}{0.9608} = 0.1411.$

24.6 The multiplication rule

24.6.1 Multiplication rule for independent processes

Consider rolling two dice. If $1/6$ of the time the first die is a 1, and $1/6$ of *those* times the second die is also a 1, then:

$$P(\text{both are 1}) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Multiplication rule for independent processes. If A and B represent events from two different and independent processes, then:

$$P(A \text{ and } B) = P(A) \times P(B)$$

If there are k events $A_1, \dots, A_k$ from k independent processes, then:

$$P(A_1 \text{ and } A_2 \text{ and } \cdots \text{ and } A_k) = P(A_1) \times P(A_2) \times \cdots \times P(A_k)$$

About 9% of people are left-handed. Suppose 2 people are selected at random from the US population. (a) What is the probability that both are left-handed? (b) What is the probability that both are right-handed?

Show answer

(a) $0.09 \times 0.09 = 0.0081$. (b) $P(\text{right-handed}) = 1 - 0.09 = 0.91$, so $0.91 \times 0.91 = 0.8281$.

Suppose 5 people are selected at random. (a) What is the probability that all are right-handed? (b) What is the probability that not all are right-handed?

Show answer

(a) $0.91^5 = 0.624$. (b) Using the complement: $P(\text{not all right-handed}) = 1 - 0.624 = 0.376$.

24.6.2 The general multiplication rule

The general multiplication rule handles events that may not be independent:

General multiplication rule. If A and B represent two outcomes or events, then:

$$P(A \text{ and } B) = P(A \mid B) \times P(B)$$

This is simply a rearrangement of the conditional probability formula.

In the smallpox data, 96.08% of residents were not inoculated, and 85.88% of those who were not inoculated survived. What is the probability that a resident was not inoculated and lived?

$$P(\text{lived and not inoculated}) = P(\text{lived} \mid \text{not inoculated}) \times P(\text{not inoculated}) = 0.8588 \times 0.9608 = 0.8251$$

24.6.3 Tree diagrams

Tree diagrams are a tool to organize outcomes and probabilities when two or more processes occur in sequence and each process is conditioned on its predecessors.

Suppose 13% of students earned an A on the midterm. Of those who earned an A on the midterm, 47% received an A on the final. Of those who earned lower than an A on the midterm, 11% received an A on the final. You randomly pick up a final exam and notice the student received an A. What is the probability the student also earned an A on the midterm?

We organize the information into a tree diagram:

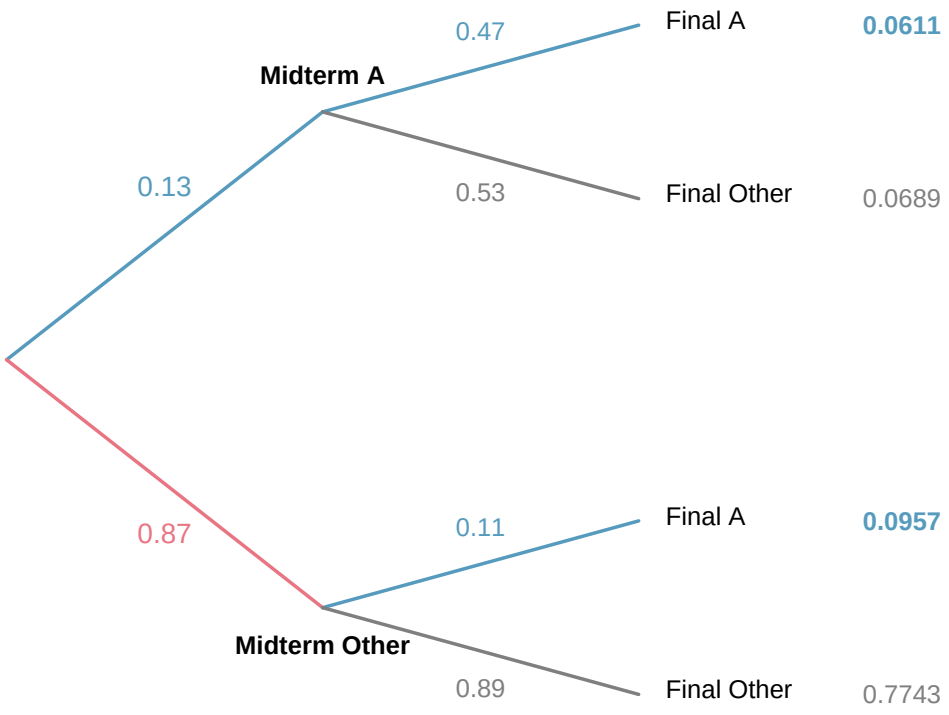


Figure 24.4: Tree diagram for midterm and final exam grades.

From the tree:

- $P(\text{midterm A and final A}) = 0.13 \times 0.47 = 0.0611$
- $P(\text{final A}) = 0.0611 + 0.0957 = 0.1568$
- $P(\text{midterm A} \mid \text{final A}) = \frac{0.0611}{0.1568} = 0.390$

The probability the student also earned an A on the midterm is about 0.39.

Sum of conditional probabilities. If $A_1, \dots, A_k$ represent all disjoint outcomes of a variable, and B is an event, then:

$$P(A_1 \mid B) + P(A_2 \mid B) + \cdots + P(A_k \mid B) = 1$$

The complement rule also holds for conditional probabilities: $P(A \mid B) = 1 - P(A^c \mid B)$.

24.7 Independence

Two processes are **independent** if knowing the outcome of one provides no useful information about the outcome of the other. Mathematically, events A and B are independent if:

$$P(A \text{ and } B) = P(A) \times P(B)$$

Equivalently, A and B are independent if $P(A \mid B) = P(A)$.

If we draw one card from a shuffled deck, is the event that the card is a heart independent of the event that the card is an ace?

$P(\heartsuit) = 1/4$, $P(\text{ace}) = 1/13$, and $P(\heartsuit \text{ and ace}) = 1/52$. We check:

$$P(\heartsuit) \times P(\text{ace}) = \frac{1}{4} \times \frac{1}{13} = \frac{1}{52} = P(\heartsuit \text{ and ace})$$

Because the equation holds, "heart" and "ace" are independent events.

Let X and Y represent the outcomes of rolling two dice. (a) What is $P(X = 1)$? (b) What is $P(Y = 1 \text{ and } X = 1)$? (c) Compute $P(Y = 1 \mid X = 1)$. (d) Compare $P(Y = 1)$ with the result from (c).

Show answer

(a) $1/6$. (b) $1/36$. (c) $\frac{1/36}{1/6} = 1/6$. (d) $P(Y = 1) = 1/6$, the same as part (c). Knowledge about X does not change the probability for Y, confirming the dice are independent.

The gambler's fallacy. Ron is watching a roulette table and notices the last five spins were all black. He bets on red, reasoning that "black six times in a row" is unlikely. This reasoning is flawed: each spin is independent of the previous spins. Past outcomes provide no information about future ones. Casinos often post recent outcomes precisely to encourage this fallacy.

24.8 Bayes' theorem (optional)

In many situations, we are given $P(B \mid A)$ but need $P(A \mid B)$ — the "inverted" conditional probability. Bayes' theorem provides a systematic way to perform this inversion.

In Canada, about 0.35% of women over 40 develop breast cancer in any given year. Mammograms produce false negatives 11% of the time (indicating no cancer when cancer is present) and false positives 7% of the time (indicating cancer when none is present). If a randomly selected woman over 40 tests positive, what is the probability she actually has breast cancer?

We organize the information into a tree diagram:

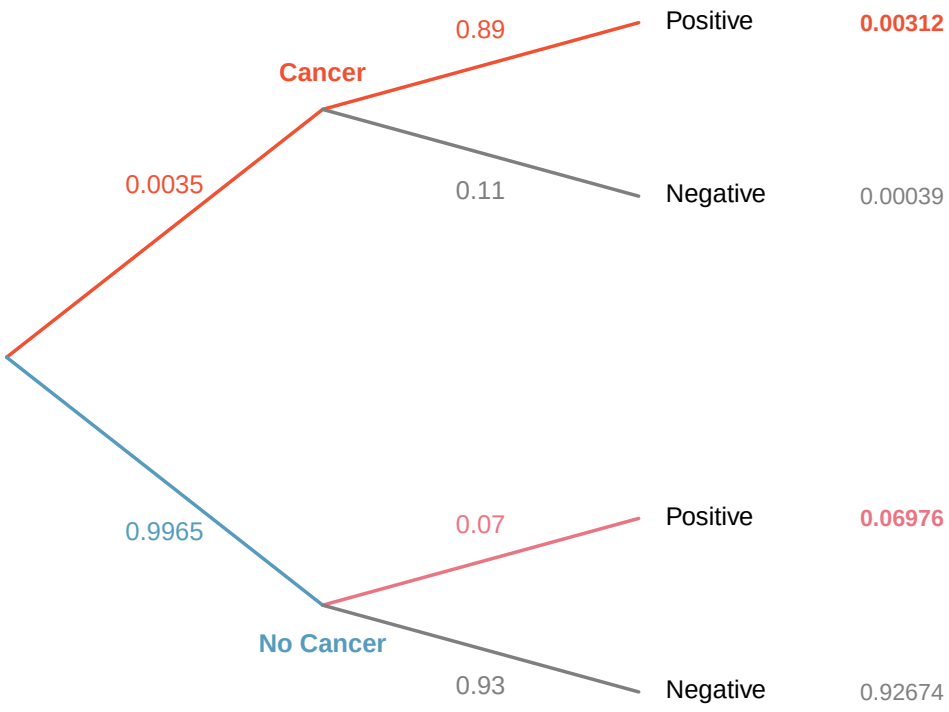


Figure 24.5: Tree diagram for breast cancer screening.

From the tree:

- $P(\text{cancer and positive}) = 0.0035 \times 0.89 = 0.00312$
- $P(\text{positive}) = 0.00312 + 0.06976 = 0.07288$
- $P(\text{cancer} \mid \text{positive}) = \frac{0.00312}{0.07288} \approx 0.043$

Even with a positive mammogram, there is only about a 4% chance the patient has breast cancer. This is why doctors often run additional tests.

Bayes' theorem. If $A_1, A_2, \ldots, A_k$ represent all possible disjoint outcomes of a first variable, and B is an observed outcome of a second variable, then:

$$P(A_1 \mid B) = \frac{P(B \mid A_1)\,P(A_1)}{P(B \mid A_1)\,P(A_1) + P(B \mid A_2)\,P(A_2) + \cdots + P(B \mid A_k)\,P(A_k)}$$

The numerator is the probability of getting both A_1 and B. The denominator is the total (marginal) probability of B.

To apply Bayes' theorem:

1. Identify the marginal probabilities: $P(A_1), P(A_2), \ldots, P(A_k)$.
2. Identify the conditional probabilities: $P(B \mid A_1), P(B \mid A_2), \ldots, P(B \mid A_k)$.
3. Substitute into the formula.

Jose visits campus on Thursday evenings. Academic events occur 35% of evenings, sporting events 20%, and no events 45%. The parking garage fills up 25% of the time during academic events, 70%

during sporting events, and 5% when there are no events. If Jose finds the garage full, what is the probability there is a sporting event?

Show answer

Using Bayes' theorem with A_1 = sporting, A_2 = academic, A_3 = none, and B = garage full: $P(\text{sporting} \mid \text{full}) = \frac{0.7 \times 0.2}{0.7 \times 0.2 + 0.25 \times 0.35 + 0.05 \times 0.45} = \frac{0.14}{0.25} = 0.56$. There is a 56% probability of a sporting event.

The strategy of updating beliefs based on new evidence is the foundation of **Bayesian statistics**, an important branch of statistics that we will not have time to explore further in this course.

24.9 Chapter review

24.9.1 Summary

In this chapter, we developed the mathematical rules of probability:

- **Probability** is the long-run proportion of times an outcome occurs, governed by the **Law of Large Numbers**.
- The **sample space** S is the set of all possible outcomes; an **event** is a subset of S.
- **Complement rule**: $P(A) = 1 - P(A^c)$.
- **Addition rule**: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. For disjoint events, $P(A \text{ and } B) = 0$.
- **Conditional probability**: $P(A \mid B) = P(A \text{ and } B)/P(B)$.
- **Multiplication rule**: $P(A \text{ and } B) = P(A \mid B) \times P(B)$. For independent events, $P(A \text{ and } B) = P(A) \times P(B)$.
- Events are **independent** if knowing one occurred does not change the probability of the other.
- **Bayes' theorem** inverts conditional probabilities.

24.10 Exercises

1. **True or false.** For each statement, decide whether it is true or false and explain your reasoning.

a. A basketball player has missed her last 4 free throws (and usually makes about 70%). Her next attempt is therefore more likely to go in than usual.

b. When rolling a fair six-sided die, the events "roll a 4" and "roll an even number" are mutually exclusive.

c. For a given Saturday, $P(\text{it rains}) + P(\text{it does not rain}) = 1$.

2. **Roulette wheel.** The game of roulette involves spinning a wheel with 38 slots: 18 red, 18 black, and 2 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball.

a. You watch a roulette wheel spin 3 consecutive times and the ball lands on a red slot each time. What is the probability that the ball will land on a red slot on the next spin?

b. You watch a roulette wheel spin 300 consecutive times and the ball lands on a red slot each time. What is the probability that the ball will land on a red slot on the next spin?

c. Are you equally confident of your answers to parts (a) and (b)? Why or why not?

3. **Four games, one winner.** Below are four versions of the same game. Your archnemesis gets to pick the version of the game, and then you get to choose how many times to flip a coin: 10 times or 100 times. Identify how many coin flips you should choose for each version of the game. It costs $1 to play each game. Explain your reasoning.

a. If the proportion of heads is larger than 0.60, you win $1.

b. If the proportion of heads is larger than 0.40, you win $1.

c. If the proportion of heads is between 0.40 and 0.60, you win $1.

d. If the proportion of heads is smaller than 0.30, you win $1.

4. **Backgammon.** Backgammon is a board game for two players in which the playing pieces are moved according to the roll of two dice. Players win by removing all of their pieces from the board, so it is usually good to roll high numbers. You are playing backgammon with a friend and you roll two 6s in your first roll and two 6s in your second roll. Your friend rolls two 3s in his first roll and again in his second row. Your friend claims that you are cheating, because rolling double 6s twice in a row is very unlikely. Using probability, show that your rolls were just as likely as his.

5. **Coin flips.** If you flip a fair coin 10 times, what is the probability of

a. getting all tails?

b. getting all heads?

c. getting at least one tails?

6. **Dice rolls.** If you roll a pair of fair dice, what is the probability of

a. getting a sum of 1?

b. getting a sum of 5?

c. getting a sum of 12?

7. **Swing voters.** A Pew Research survey asked 2,373 randomly sampled registered voters their political affiliation (Republican, Democrat, or Independent) and whether or not they identify as swing voters. 35% of respondents identified as Independent, 23% identified as swing voters, and 11% identified as both.

a. Are being Independent and being a swing voter disjoint, i.e. mutually exclusive?

b. Draw a Venn diagram summarizing the variables and their associated probabilities.

c. What percent of voters are Independent but not swing voters?

d. What percent of voters are Independent or swing voters?

e. What percent of voters are neither Independent nor swing voters?

f. Is the event that someone is a swing voter independent of the event that someone is a political Independent?

8. **Poverty and language.** The American Community Survey is an ongoing survey that provides data every year to give communities the current information they need to plan investments and services. The 2010 American Community Survey estimates that 14.6% of Americans live below the poverty line, 20.7% speak a language other than English (foreign language) at home, and 4.2% fall into both categories.

a. Are living below the poverty line and speaking a foreign language at home disjoint?

b. Draw a Venn diagram summarizing the variables and their associated probabilities.

c. What percent of Americans live below the poverty line and only speak English at home?

d. What percent of Americans live below the poverty line or speak a foreign language at home?

e. What percent of Americans live above the poverty line and only speak English at home?

f. Is the event that someone lives below the poverty line independent of the event that the person speaks a foreign language at home?

9. **Disjoint vs. independent.** In parts (a) and (b), identify whether the events are disjoint, independent, or neither (events cannot be both disjoint and independent).

a. You and a randomly selected student from your class both earn A's in this course.

b. You and your class study partner both earn A's in this course.

c. If two events can occur at the same time, must they be dependent?

10. **Guessing on an exam.** In a multiple choice exam, there are 5 questions and 4 choices for each question (a, b, c, d). Nancy has not studied for the exam at all and decides to randomly guess the answers. What is the probability that:

a. the first question she gets right is the 5th question?

b. she gets all of the questions right?

c. she gets at least one question right?

11. **Educational attainment of couples.** The table below shows the distribution of education level attained by US residents by gender based on data collected in the 2010 American Community Survey.

c c }		*Gender*
	9th to 12th grade, no diploma	0.10
Highest	HS graduate (or equivalent)	0.30
education	Some college, no degree	0.22
attained	Associate's degree	0.06
	Bachelor's degree	0.16
	Graduate or professional degree	0.09

a. What is the probability that a randomly chosen man has at least a Bachelor's degree?

b. What is the probability that a randomly chosen woman has at least a Bachelor's degree?

c. What is the probability that a man and a woman getting married both have at least a Bachelor's degree? Note any assumptions you must make to answer this question.

d. If you made an assumption in part (c), do you think it was reasonable? If you didn't make an assumption, double check your earlier answer and then return to this part.

12. **School absences.** Data collected at elementary schools in DeKalb County, GA suggest that each year roughly 25% of students miss exactly one day of school, 15% miss 2 days, and 28% miss 3 or more days due to sickness.

a. What is the probability that a student chosen at random doesn't miss any days of school due to sickness this year?

b. What is the probability that a student chosen at random misses no more than one day?

c. What is the probability that a student chosen at random misses at least one day?

d. If a parent has two kids at a DeKalb County elementary school, what is the probability that neither kid will miss any school? Note any assumption you must make to answer this question.

e. If a parent has two kids at a DeKalb County elementary school, what is the probability that both kids will miss some school, i.e. at least one day? Note any assumption you make.

f. If you made an assumption in part (d) or (e), do you think it was reasonable? If you didn't make any assumptions, double check your earlier answers.

13. **Joint and conditional probabilities.** $P(A) = 0.3$, $P(B) = 0.7$

a. Can you compute $P(A \text{ and } B)$ if you only know $P(A)$ and $P(B)$?

b. Assuming that events A and B arise from independent random processes,

c. what is $P(A \text{ and } B)$?

d. what is $P(A \text{ or } B)$?

e. what is $P(A|B)$?

f. If we are given that $P(A \text{ and } B) = 0.1$, are the random variables giving rise to events A and B independent?

g. If we are given that $P(A \text{ and } B) = 0.1$, what is $P(A|B)$?

14. **PB & J.** Suppose 80% of people like peanut butter, 89% like jelly, and 78% like both. Given that a randomly sampled person likes peanut butter, what's the probability that he also likes jelly?

15. **Global warming.** A Pew Research poll asked 1,306 Americans "From what you've read and heard, is there solid evidence that the average temperature on earth has been getting warmer over the past few decades, or not?". The table below shows the distribution of responses by party and ideology, where the counts have been replaced with relative frequencies.

		Response
		warming
Party and Ideology	Mod/Lib Republican	0.06
	Mod/Cons Democrat	0.25
	Liberal Democrat	0.18

a. Are believing that the earth is warming and being a liberal Democrat mutually exclusive?

b. What is the probability that a randomly chosen respondent believes the earth is warming or is a liberal Democrat?

c. What is the probability that a randomly chosen respondent believes the earth is warming given that he is a liberal Democrat?

d. What is the probability that a randomly chosen respondent believes the earth is warming given that he is a conservative Republican?

e. Does it appear that whether or not a respondent believes the earth is warming is independent of their party and ideology? Explain your reasoning.

f. What is the probability that a randomly chosen respondent is a moderate/liberal Republican given that he does not believe that the earth is warming?

16. **Health coverage, relative frequencies.** The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table displays the distribution of health status of respondents to this survey (excellent, very good, good, fair, poor) and whether or not they have health insurance.

		Health Status	
Coverage	Yes	0.2099	0.3123

a. Are being in excellent health and having health coverage mutually exclusive?

b. What is the probability that a randomly chosen individual has excellent health?

c. What is the probability that a randomly chosen individual has excellent health given that he has health coverage?

d. What is the probability that a randomly chosen individual has excellent health given that he doesn't have health coverage?

e. Do having excellent health and having health coverage appear to be independent?

17. **Burger preferences.** A 2010 SurveyUSA poll asked 500 Los Angeles residents, "What is the best hamburger place in Southern California? Five Guys Burgers? In-N-Out Burger? Fat Burger? Tommy's Hamburgers? Umami Burger? Or somewhere else?" The distribution of responses by gender is shown below.

r r r }		*Gender*
	In-N-Out Burger	162
Best	Fat Burger	10
hamburger	Tommy's Hamburgers	27
place	Umami Burger	5
	Other	26
	Not Sure	13

a. Are being female and liking Five Guys Burgers mutually exclusive?

b. What is the probability that a randomly chosen male likes In-N-Out the best?

c. What is the probability that a randomly chosen female likes In-N-Out the best?

d. What is the probability that a man and a woman who are dating both like In-N-Out the best? Note any assumption you make and evaluate whether you think that assumption is reasonable.

e. What is the probability that a randomly chosen person likes Umami best or that person is female?

18. **Assortative mating.** Assortative mating is a nonrandom mating pattern where individuals with similar genotypes and/or phenotypes mate with one another more frequently than what would be expected under a random mating pattern. Researchers studying this topic collected data on eye colors of 204 Scandinavian men and their female partners. The table below summarizes the results.

		Partner (*female*)

		Partner (*female*)
Self (*male*)	Brown	19
	Green	11

a. What is the probability that a randomly chosen male respondent or his partner has blue eyes?

b. What is the probability that a randomly chosen male respondent with blue eyes has a partner with blue eyes?

c. What is the probability that a randomly chosen male respondent with brown eyes has a partner with blue eyes? What about the probability of a randomly chosen male respondent with green eyes having a partner with blue eyes?

d. Does it appear that the eye colors of male respondents and their partners are independent? Explain your reasoning.

19. **Tutoring and passing.** A campus tutoring center reports that 70% of statistics students attend tutoring at least once during the semester. Of students who attend, 90% pass the course; of students who don't attend, 72% pass.

a. Find the probability that a randomly selected student both attends tutoring AND passes.

b. Find the overall probability that a randomly selected student passes.

c. Given that a student passed, find the probability they attended tutoring.

20. **Counterfeit detector accuracy.** A bank's automated bill-handling machine scans bills as they are deposited. About 3% of bills it processes are counterfeit. The scanner correctly flags 97% of counterfeit bills as suspicious (its sensitivity), and it has a 5% false-positive rate on genuine bills.

a. What is the overall probability that a randomly chosen bill is flagged as suspicious?

b. Given that a bill is flagged, what is the probability it is actually counterfeit?

c. Explain in one sentence why your answer to (b) is much smaller than the scanner's sensitivity.

21. **It's never lupus.** Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to a high risk of blood clotting. It is believed that 2% of the population suffer from this disease. The test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease. There is a line from the Fox television show *House* that is often used after a patient tests positive for lupus: "It's never lupus." Do you think there is truth to this statement? Use appropriate probabilities to support your answer.

22. **Exit poll.** Edison Research gathered exit poll results from several sources for the Wisconsin recall election of Scott Walker. They found that 53% of the respondents voted in favor of Scott Walker. Additionally, they estimated that of those who did vote in favor for Scott Walker, 37% had a college degree, while 44% of those who voted against Scott Walker had a college degree. Suppose we randomly sampled a person who participated in the exit poll and found that he had a college degree. What is the probability that he voted in favor of Scott Walker?

23. **Marbles in an urn.** Imagine you have an urn containing 5 red, 3 blue, and 2 orange marbles in it.

a. What is the probability that the first marble you draw is blue?

b. Suppose you drew a blue marble in the first draw. If drawing with replacement, what is the probability of drawing a blue marble in the second draw?

c. Suppose you instead drew an orange marble in the first draw. If drawing with replacement, what is the probability of drawing a blue marble in the second draw?

d. If drawing with replacement, what is the probability of drawing two blue marbles in a row?

e. When drawing with replacement, are the draws independent? Explain.

24. **Socks in a drawer.** In your sock drawer you have 4 blue, 5 gray, and 3 black socks. Half asleep one morning you grab 2 socks at random and put them on. Find the probability you end up wearing

a. 2 blue socks

b. no gray socks

c. at least 1 black sock

d. a green sock

e. matching socks

25. **Chips in a bag.** Imagine you have a bag containing 5 red, 3 blue, and 2 orange chips.

a. Suppose you draw a chip and it is blue. If drawing without replacement, what is the probability the next is also blue?

b. Suppose you draw a chip and it is orange, and then you draw a second chip without replacement. What is the probability this second chip is blue?

c. If drawing without replacement, what is the probability of drawing two blue chips in a row?

d. When drawing without replacement, are the draws independent? Explain.

26. **Books on a bookshelf.** The table below shows the distribution of books on a bookcase based on whether they are nonfiction or fiction and hardcover or paperback.

	Format
Nonfiction	15

a. Find the probability of drawing a hardcover book first then a paperback fiction book second when drawing without replacement.

b. Determine the probability of drawing a fiction book first and then a hardcover book second, when drawing without replacement.

c. Calculate the probability of the scenario in part (b), except this time complete the calculations under the scenario where the first book is placed back on the bookcase before randomly drawing the second book.

d. The final answers to parts (b) and (c) are very similar. Explain why this is the case.

27. **Student outfits.** In a classroom with 24 students, 7 students are wearing jeans, 4 are wearing shorts, 8 are wearing skirts, and the rest are wearing leggings. If we randomly select 3 students without replacement, what is the probability that one of the selected students is wearing leggings and the other two are wearing jeans? Note that these are mutually exclusive clothing options.

28. **The birthday problem.** Suppose we pick three people at random. For each of the following questions, ignore the special case where someone might be born on February 29th, and assume that births are evenly distributed throughout the year.

a. What is the probability that the first two people share a birthday?

b. What is the probability that at least two people share a birthday?

29. **Health coverage, frequencies.** The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table summarizes two variables for the respondents: health status and health coverage, which describes whether each respondent had health insurance.

		Health Status	
Coverage	Yes	4,198	6,245

a. If we draw one individual at random, what is the probability that the respondent has excellent health and doesn't have health coverage?

b. If we draw one individual at random, what is the probability that the respondent has excellent health or doesn't have health coverage?

30. **HIV in Swaziland.** Swaziland has the highest HIV prevalence in the world: 25.9% of this country's population is infected with HIV. The ELISA test is one of the first and most accurate tests for HIV. For those who carry HIV, the ELISA test is 99.7% accurate. For those who do not carry HIV, the test is 92.6% accurate. If an individual from Swaziland has tested positive, what is the probability that he carries HIV?

31. **Defective sample from one machine.** A factory produces parts on two machines. Machine A makes 40% of the output and has a 2% defect rate; Machine B makes the remaining 60% with a 5% defect rate. You take three parts from the same (unknown) machine and find that all three are defective.

a. Given that all three parts are defective, what is the probability they came from Machine A?

b. Why is the answer in (a) much smaller than the prior probability that a randomly chosen machine is Machine A?

Chapter 25

Random Variables

Extended Content. This chapter extends beyond a standard one-semester introductory course. It is included for completeness and for instructors who wish to cover additional topics. Content is in draft form.

In the previous chapter, we developed rules for computing probabilities of events. Now we shift our focus to **random variables** — numerical summaries of random processes. A random variable assigns a number to each outcome of a random process, which allows us to apply mathematical tools for prediction and analysis. We distinguish between discrete and continuous random variables, introduce probability distributions, and build the foundation for the specific distributions (binomial and normal) that follow in later chapters.

25.1 Random variables

It is often useful to model a random process by assigning a number to each outcome. This numerical model is called a **random variable**.

A **random variable** is a numerical outcome of a random process. We typically denote random variables with capital letters such as X, Y, or Z, and their observed values with the corresponding lowercase letters x, y, or z.

Two books are assigned for a statistics class: a textbook ($137) and a study guide ($33). The university bookstore has determined that 20% of enrolled students buy neither book, 55% buy the textbook only, and 25% buy both books. If X represents the amount a single student spends on books, then X is a random variable with possible values $0, $137, and $170.

Random variables come in two types — discrete and continuous — and the distinction matters because they require different mathematical tools.

25.2 Discrete random variables

A **discrete random variable** takes on a finite or countably infinite number of distinct values. These values can be listed (even if the list is infinitely long).

Examples of discrete random variables:

- The number showing when you roll a die: $X \in \{1, 2, 3, 4, 5, 6\}$

- The number of heads in 10 coin flips: $X \in \{0, 1, 2, \ldots, 10\}$
- The number of customers arriving at a store in an hour: $X \in \{0, 1, 2, 3, \ldots\}$

25.2.1 Why the distinction matters

In Parts II and III, we worked extensively with sample statistics like the sample mean $\bar{x}$ and the sample proportion $\hat{p}$. These are actually *observed values* of random variables. The sampling distribution of $\bar{x}$ that we studied in the sampling distributions chapter is really the probability distribution of a random variable. Formalizing the language of random variables helps us state results more precisely and connect our earlier simulation-based reasoning to mathematical theory.

25.2.2 Probability mass functions

The complete description of a discrete random variable is given by its **probability mass function** (PMF), which specifies the probability of each possible value.

The **probability mass function** (PMF) of a discrete random variable X is the function $P(X = x)$ that gives the probability that X takes on each possible value x. A valid PMF must satisfy:

1. $P(X = x) \geq 0$ for every value x.
2. The sum of all probabilities equals 1: $\sum_{\text{all } x} P(X = x) = 1$.

The bookstore example from the previous section has the following PMF:

x_i	\$0	\$137	\$170
$P(X = x_i)$	0.20	0.55	0.25

We can verify this is a valid PMF: all probabilities are between 0 and 1, and $0.20 + 0.55 + 0.25 = 1.00$.

Suppose a campus shuttle arrives every 10 minutes. The number of students who board the shuttle at a particular stop can be modeled with the following distribution:

Number of students (x)	0	1	2	3	4
$P(X = x)$	0.10	0.25	0.35	0.20	0.10

(a) Verify this is a valid probability distribution. (b) What is the probability that 2 or fewer students board? (c) What is the probability that more than 2 students board?

Show answer

(a) All probabilities are between 0 and 1, and $0.10 + 0.25 + 0.35 + 0.20 + 0.10 = 1.00$. (b) $P(X \leq 2) = 0.10 + 0.25 + 0.35 = 0.70$. (c) $P(X > 2) = 1 - P(X \leq 2) = 1 - 0.70 = 0.30$, or equivalently $0.20 + 0.10 = 0.30$.

25.2.3 Visualizing discrete distributions

Discrete probability distributions are often displayed as bar plots where the height of each bar represents the probability of that outcome. When the outcomes are numerical, we place the bars at their numerical locations, creating a plot that resembles a histogram.

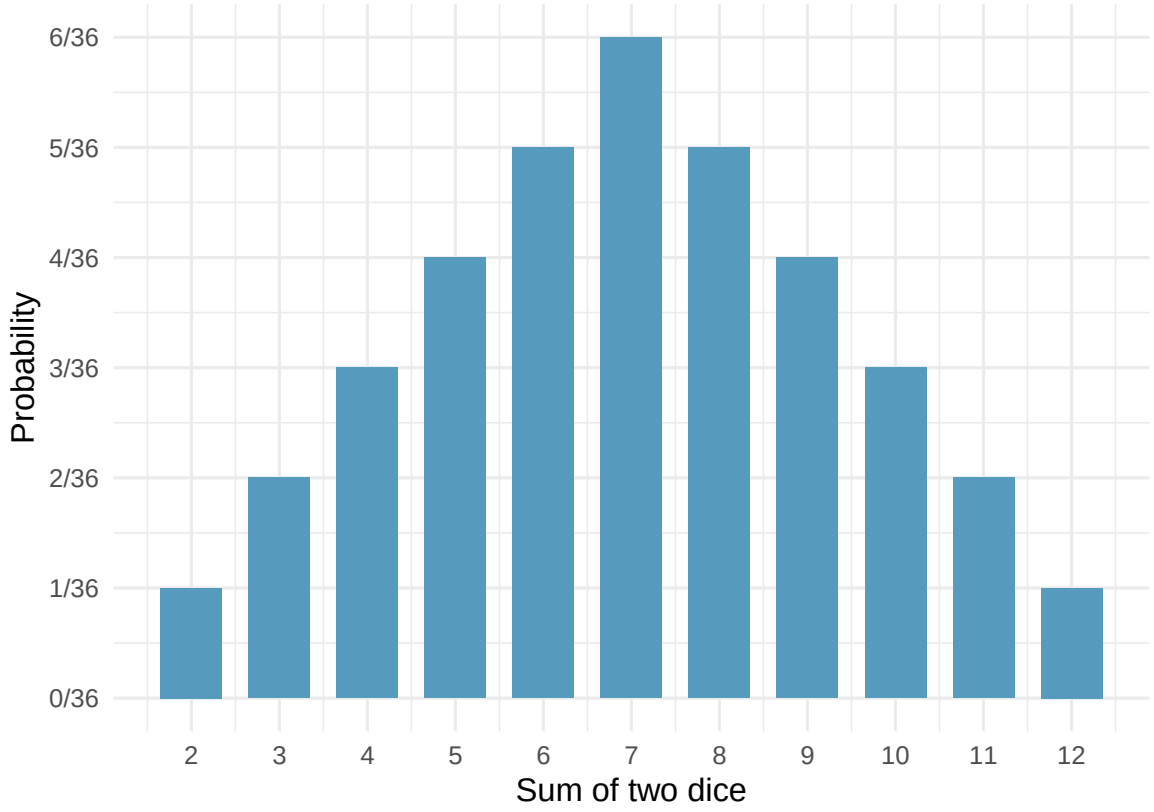


Figure 25.1: Probability distribution for the sum of two dice displayed as a bar plot.

25.2.4 Building intuition with simulation

We can also *estimate* a probability distribution by simulating the random process many times and recording the relative frequencies of each outcome. For example, if we simulate rolling two dice 10,000 times and record the sum each time, the histogram of those 10,000 sums will closely approximate the theoretical probability distribution. As the number of simulations increases, the approximation improves — this is the Law of Large Numbers in action.

This connection between simulation and theory is one of the most important ideas in the course. Every time we used simulation in Parts II and III to build a null distribution or a bootstrap distribution, we were approximating a theoretical probability distribution.

25.3 Cumulative distribution functions

The **cumulative distribution function** (CDF) of a random variable X is the function:

$$F(x) = P(X \leq x)$$

The CDF gives the probability that X takes a value less than or equal to x.

The CDF is a useful alternative to the PMF for answering "at most" or "at least" types of questions. For a discrete random variable, the CDF is a step function that jumps at each possible value of X.

Consider the sum of two dice. The CDF at a few key values:

x	2	3	4	5	6	7	8	9	10	11	12
$F(x) = P(X \leq x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

Key properties of any CDF:

- $F(x)$ is non-decreasing: if $a < b$, then $F(a) \leq F(b)$.
- $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to +\infty$.
- $P(a < X \leq b) = F(b) - F(a)$.

The CDF becomes especially important for continuous random variables, where individual values have zero probability and we must always work with intervals.

Using the CDF for the sum of two dice above, find: (a) $P(X \leq 5)$. (b) $P(X > 9)$. (c) $P(4 < X \leq 8)$.

Show answer

(a) $F(5) = 10/36$. (b) $P(X > 9) = 1 - F(9) = 1 - 30/36 = 6/36 = 1/6$. (c) $P(4 < X \leq 8) = F(8) - F(4) = 26/36 - 6/36 = 20/36 = 5/9$.

25.4 Continuous random variables

A **continuous random variable** can take any value in some interval of real numbers. Its probability distribution is described by a smooth curve called a **probability density function** (PDF).

Examples of continuous random variables:

- The height of a randomly selected adult
- The time until the next bus arrives
- The temperature at noon tomorrow

25.4.1 From histograms to density curves

When we have a very large sample and use very narrow bins in a histogram, the outline of the histogram starts to look like a smooth curve. This smooth curve is a probability density function.

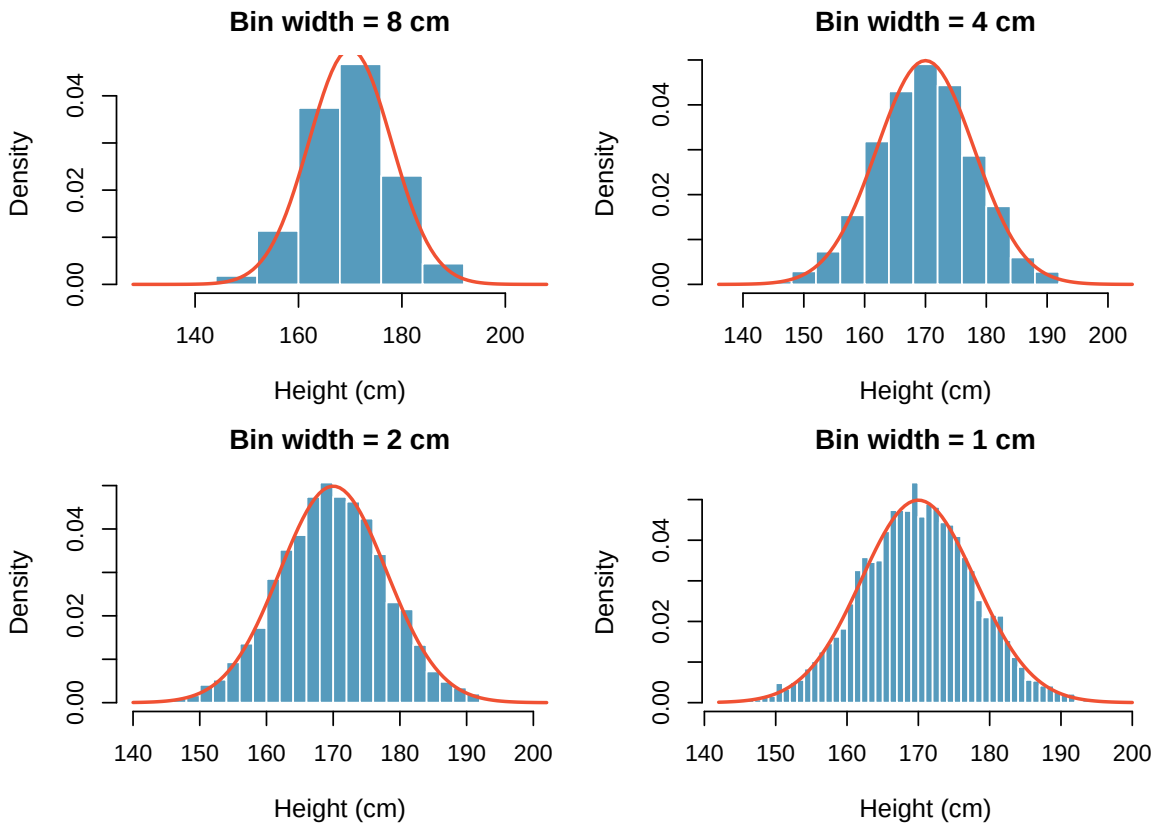


Figure 25.2: The transition from a histogram to a smooth density curve as the bin width decreases and sample size increases. ↗ Try this live — overlay a density curve on the histogram of your own data.

Properties of a probability density function (PDF):

1. The density curve is always on or above the horizontal axis: $f(x) \geq 0$ for all x.
2. The total area under the curve equals 1.
3. The probability that X falls between a and b equals the area under the curve between a and b: $P(a \leq X \leq b) =$ area between a and b.

25.4.2 Probabilities from continuous distributions

For a continuous random variable, probabilities are *areas* under the density curve.

The heights of US adults follow a continuous distribution. What proportion of the population has heights between 180 cm and 185 cm?

Using the density curve, we find the area under the curve between 180 and 185:

$$P(180 \leq X \leq 185) = \text{area between 180 and 185} \approx 0.1157$$

About 11.6% of US adults are between 180 and 185 cm tall.

A key difference: the probability of an exact value is zero for continuous random variables.

For a continuous random variable, $P(X = c) = 0$ for any specific value c. There is no area under the curve at a single point. This means:

$$P(X \leq a) = P(X < a)$$

This is *not* true for discrete random variables, where individual outcomes carry positive probability.

If $P(180 \leq \text{height} \leq 185) = 0.1157$ for a randomly selected US adult, and three adults are selected at random: (a) What is the probability all three have heights between 180 and 185 cm? (b) What is the probability none of them do?

Show answer

(a) Since the selections are independent: $0.1157^3 \approx 0.0015$. (b) $(1-0.1157)^3 = 0.8843^3 \approx 0.692$.

25.4.3 Discrete vs. continuous: a summary

Feature	Discrete	Continuous
Values	Finite or countable list	Any value in an interval
Described by	Probability mass function (PMF)	Probability density function (PDF)
$P(X = c)$	Can be positive	Always equals 0
Probabilities	Sum probabilities of individual values	Find areas under the curve
Visualization	Bar plot	Smooth curve
Example	Number of coin flips until heads	Height of a randomly selected person

In this course, we will work primarily with two specific distributions:

- The **binomial distribution** (the binomial distribution chapter) — a discrete distribution
- The **normal distribution** (the normal distribution chapter) — a continuous distribution

Understanding the general concepts from this chapter will make those specific distributions much easier to work with.

25.5 Chapter review

25.5.1 Summary

- A **random variable** assigns a numerical value to each outcome of a random process.
- **Discrete random variables** take on a finite or countable set of values. Their distributions are described by probability mass functions (PMFs).
- **Continuous random variables** can take any value in an interval. Their distributions are described by probability density functions (PDFs), and probabilities are calculated as areas under the curve.
- The **cumulative distribution function** (CDF) gives $P(X \leq x)$ and works for both discrete and continuous random variables.
- For continuous random variables, $P(X = c) = 0$ for any specific value c.
- A valid probability distribution has non-negative probabilities that sum (or integrate) to 1.

25.6 Exercises

1. **College smokers.** At a university, 13% of students smoke.

a. Calculate the expected number of smokers in a random sample of 100 students from this university.

b. The university gym opens at 9 am on Saturday mornings. One Saturday morning at 8:55 am there are 27 students outside the gym waiting for it to open. Should you use the same approach from part (a) to calculate the expected number of smokers among these 27 students?

2. **Carnival prize wheel.** A carnival prize wheel has 10 equally-likely sectors: 5 say "Try Again" (win \$0), 3 pay \$3, 1 pays \$10, and 1 pays a \$25 jackpot. Let W be your winnings from one spin.

a. Compute $E(W)$ and SD(W).

b. What is the maximum amount a risk-neutral player should be willing to pay per spin? Explain.

3. **Marble draws without replacement.** A carnival booth has a bag of 12 marbles: 5 red, 4 blue, and 3 green. You draw 3 marbles without replacement. The payouts are \$40 for 3 reds, \$25 for 3 blues, and \$0 for any other draw. Let W be your winnings.

a. Compute $E(W)$ and SD(W).

b. The booth charges \$5 to play. Let $N = W - 5$ be your net profit on a single play. Compute $E(N)$ and SD(N).

c. Based on $E(N)$, should a risk-neutral player play this game? Explain.

4. **Is it worth it.** Andy is always looking for ways to make money fast. Lately, he has been trying to make money by gambling. Here is the game he is considering playing: The game costs \$2 to play. He draws a card from a deck. If he gets a number card (2-10), he wins nothing. For any face card (jack, queen or king), he wins \$3. For any ace, he wins \$5, and he wins an *extra* \$20 if he draws the ace of clubs.

a. Create a probability model and find Andy's expected profit per game.

b. Would you recommend this game to Andy as a good way to make money? Explain.

5. **Cat weights.** The histogram shown below represents the weights (in kg) of 47 female and 97 male cats.

a. What fraction of these cats weigh less than 2.5 kg?

b. What fraction of these cats weigh between 2.5 and 2.75 kg?

c. What fraction of these cats weigh between 2.75 and 3.5 kg?

6. **Income and gender.** The relative frequency table below displays the distribution of annual total personal income (in 2009 inflation-adjusted dollars) for a representative sample of 96,420,486 Americans. These data come from the American Community Survey for 2005-2009. This sample is comprised of 59% males and 41% females.

a. Describe the distribution of total personal income.

b. What is the probability that a randomly chosen US resident makes less than \$50,000 per year?

c. What is the probability that a randomly chosen US resident makes less than \$50,000 per year and is female? Note any assumptions you make.

d. The same data source indicates that 71.8% of females make less than \$50,000 per year. Use this value to determine whether or not the assumption you made in part (c) is valid.

{

$10,000 to $14,999	4.7%
$15,000 to $24,999	15.8%
$25,000 to $34,999	18.3%
$35,000 to $49,999	21.2%
$50,000 to $64,999	13.9%
$65,000 to $74,999	5.8%
$75,000 to $99,999	8.4%
$100,000 or more	9.7%

}

7. **Grade distributions.** Each row in the table below is a proposed grade distribution for a class. Identify each as a valid or invalid probability distribution, and explain your reasoning.

	Grades
(b)	0
(c)	0.3
(d)	0.3
(e)	0.2
(f)	0

8. **Three-round trivia bet.** A trivia contestant has a 50% chance of answering any given question correctly, and questions are independent. She plays three rounds. Each round she places a $1 bet: a correct answer wins her $1, a wrong answer loses $1. Let Y be her total winnings after the three rounds.

The possible values of Y are $-\$3, -\$1, +\$1, +\$3$. Compute the probability of each, and then compute $E(Y)$.

Chapter 26

Expected Value and Variance

Extended Content. This chapter extends beyond a standard one-semester introductory course. It is included for completeness and for instructors who wish to cover additional topics. Content is in draft form.

How much revenue should a bookstore expect per student? How much variability should a stock investor anticipate in their portfolio? These questions require us to summarize a random variable's distribution with just two numbers: its **expected value** (center) and **variance** (spread). In this chapter, we define these quantities, develop useful computational properties, and extend them to linear combinations of random variables. These tools will be essential when we study the binomial and normal distributions in the chapters that follow.

26.1 Expected value of a random variable

We introduced random variables in the random variables chapter. Now we develop the tools to summarize them numerically, starting with the expected value — the long-run average of a random variable.

Two books are assigned for a statistics class: a textbook ($137) and a study guide ($33). From past experience, the bookstore knows that 20% of enrolled students buy neither book, 55% buy the textbook only, and 25% buy both books. If there are 100 students enrolled, how many books should the bookstore expect to sell?

About 20 students buy 0 books, 55 buy 1 book (55 books total), and 25 buy 2 books (50 books total). The bookstore should expect to sell about $0 + 55 + 50 = 105$ books.

What is the average revenue per student?

The expected total revenue is $\$0 \times 20 + \$137 \times 55 + \$170 \times 25 = \$11{,}785$, and there are 100 students. The expected revenue per student is $\$11{,}785/100 = \$117.85$.

Notice that the "average revenue per student" was computed by multiplying each outcome by its probability and summing:

$$0 \times 0.20 + 137 \times 0.55 + 170 \times 0.25 = 117.85$$

This is the **expected value** of the random variable X.

Expected value of a discrete random variable. If X takes outcomes $x_1, x_2, \ldots, x_k$ with probabilities $P(X = x_1), P(X = x_2), \ldots, P(X = x_k)$, the **expected value** of X is:

$$E(X) = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \cdots + x_k \cdot P(X = x_k) = \sum_{i=1}^{k} x_i \, P(X = x_i)$$

The Greek letter μ may be used in place of $E(X)$.

The expected value represents the **long-run average** of the random variable. If we were to observe the random process many, many times, the average of all observed values would converge to $E(X)$.

26.1.1 The expected value as a center of gravity

In physics, the expected value corresponds to the **center of gravity** (or balance point) of a distribution. Imagine cutting out the shape of a probability distribution from a piece of cardboard. The expected value is the point where you could balance the cutout on the tip of your finger. This is why we sometimes call $E(X)$ the "mean" — it is exactly analogous to the sample mean $\bar{x}$, but for a probability distribution rather than a data set.

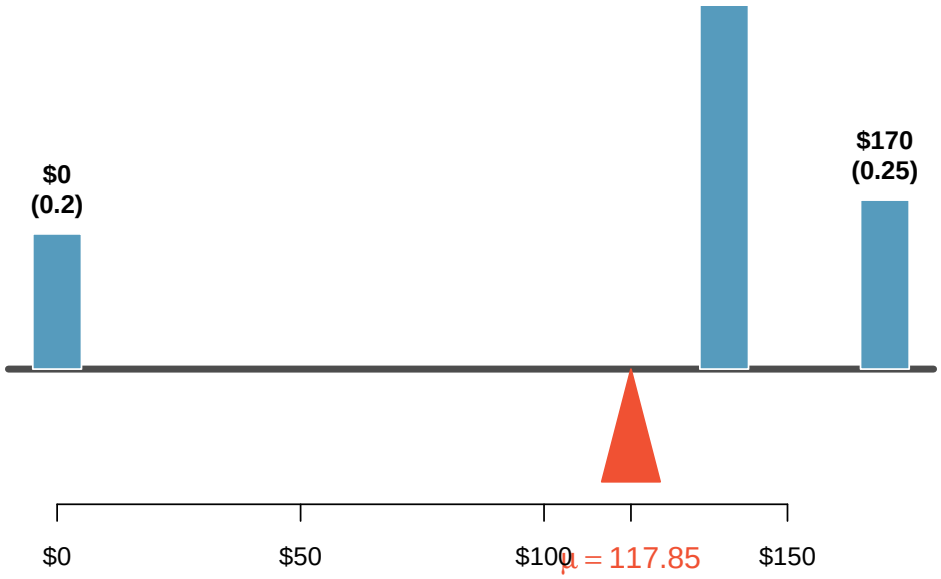


Figure 26.1: A weight system representing the bookstore distribution balanced at the mean of $117.85.

A campus shuttle arrives every 10 minutes. The number of students boarding follows this distribution:

x	0	1	2	3	4
$P(X = x)$	0.10	0.25	0.35	0.20	0.10

What is the expected number of students who board?

Show answer

$E(X) = 0(0.10) + 1(0.25) + 2(0.35) + 3(0.20) + 4(0.10) = 0 + 0.25 + 0.70 + 0.60 + 0.40 = 1.95$. On average, about 1.95 students board the shuttle at this stop.

The expected value does not have to be a value that the random variable can actually take. In the shuttle example, $E(X) = 1.95$, but you can never have 1.95 students board. The expected value is a long-run average, not a prediction of any single outcome.

26.2 Variance of a random variable

The expected value tells us where the center of a distribution lies, but it says nothing about the *spread*. The bookstore might want to know not just the average revenue per student, but how much that revenue varies from student to student.

Variance of a discrete random variable. If X takes outcomes $x_1, \ldots, x_k$ with probabilities $P(X = x_1), \ldots, P(X = x_k)$ and expected value $\mu = E(X)$, then the **variance** of X is:

$$\text{Var}(X) = \sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 \, P(X = x_i)$$

The **standard deviation** is $\sigma = \sqrt{\text{Var}(X)}$.

The variance is the average squared deviation from the mean, weighted by the probabilities. The standard deviation has the same units as X, which makes it more interpretable.

Compute the variance and standard deviation of the bookstore revenue per student.

We organize the calculation in a table with $\mu = 117.85$:

	$x_1 = 0$	$x_2 = 137$	$x_3 = 170$	Total
$P(X = x_i)$	0.20	0.55	0.25	
$x_i - \mu$	-117.85	19.15	52.15	
$(x_i - \mu)^2$	13,888.6	366.7	2,719.6	
$(x_i - \mu)^2 \cdot P(X = x_i)$	2,777.7	201.7	679.9	3,659.3

The variance is $\sigma^2 = 3{,}659.3$ and the standard deviation is $\sigma = \sqrt{3{,}659.3} = \$60.49$.

The bookstore also sells a chemistry textbook ($159) and supplement ($41). From past experience, 15% of chemistry students buy neither, 25% buy the textbook only, and 60% buy both. Let Y be the revenue from a single chemistry student. (a) Write out the probability distribution of Y. (b) Compute $E(Y)$. (c) Compute SD(Y).

Show answer

(a) Y takes values \$0, \$159, \$200 with probabilities 0.15, 0.25, 0.60. (b) $E(Y) = 0(0.15) + 159(0.25) + 200(0.60) = 0 + 39.75 + 120 = \$159.75$. (c) $\text{Var}(Y) = (0 - 159.75)^2(0.15) + (159 - 159.75)^2(0.25) + (200 - 159.75)^2(0.60) = 3828.0 + 0.1 + 972.0 = 4800.1$, so SD$(Y) = \sqrt{4800.1} \approx \$69.28$.

26.3 Properties of expected value

The expected value has several important algebraic properties that make it easy to work with. These properties follow from the definition but are worth stating explicitly because we will use them frequently.

Properties of expected value. For a random variable X and constants a and b:

1. **Adding a constant shifts the mean**: $E(X + b) = E(X) + b$
2. **Multiplying by a constant scales the mean**: $E(aX) = a \cdot E(X)$
3. **General linear transformation**: $E(aX + b) = a \cdot E(X) + b$

Why these make sense: If every student's textbook cost increases by \$10 (adding a constant $b = 10$), then the average cost also increases by \$10. If the bookstore marks up all prices by 20% (multiplying by $a = 1.20$), then the average revenue also increases by 20%.

Suppose the bookstore offers a 10% discount on every book purchase. The discounted price is $0.90 \times X$, where X is the original price. What is the expected revenue per student under the discount?

Using the property $E(aX) = aE(X)$:

$$E(0.90X) = 0.90 \times E(X) = 0.90 \times 117.85 = \$106.07$$

The expected revenue per student drops from \$117.85 to \$106.07.

The bookstore charges a flat \$5 processing fee in addition to book costs. If X is the cost of books and the total charge is $Y = X + 5$, what is $E(Y)$?

Show answer

$E(Y) = E(X + 5) = E(X) + 5 = 117.85 + 5 = \$122.85$.

26.4 Properties of variance

Variance has different algebraic properties than expected value. The most important distinction: adding a constant does not change the variance.

Properties of variance. For a random variable X and constants a and b:

1. **Adding a constant does not change variance**: $\text{Var}(X + b) = \text{Var}(X)$
2. **Multiplying by a constant scales variance by the square**: $\text{Var}(aX) = a^2 \cdot \text{Var}(X)$
3. **General linear transformation**: $\text{Var}(aX + b) = a^2 \cdot \text{Var}(X)$

For standard deviations:

$$\text{SD}(aX + b) = |a| \cdot \text{SD}(X)$$

Why adding a constant does not change variance: If every student pays \$5 more, the *spread* of payments does not change — everyone's payment shifts by the same amount. The center moves, but the variability stays the same.

Why multiplying scales by the square: If every price is doubled ($a = 2$), then deviations from the mean are also doubled, so *squared* deviations are multiplied by $2^2 = 4$.

For the bookstore revenue with $\text{Var}(X) = 3{,}659.3$ and $\sigma_X = \$60.49$:

- If a \$5 fee is added: $\text{Var}(X + 5) = \text{Var}(X) = 3{,}659.3$ and $\text{SD}(X + 5) = \$60.49$.
- If a 10% discount is applied: $\text{Var}(0.90X) = 0.90^2 \times 3{,}659.3 = 0.81 \times 3{,}659.3 = 2{,}964.0$ and $\text{SD}(0.90X) = 0.90 \times 60.49 = \$54.44$.

Suppose temperatures in Celsius are converted to Fahrenheit using $F = 1.8C + 32$. If the standard deviation of daily high temperatures is $\sigma_C = 5°\text{C}$, what is the standard deviation in Fahrenheit?

Show answer

$\text{SD}(1.8C + 32) = 1.8 \times \text{SD}(C) = 1.8 \times 5 = 9°\text{F}$. The additive constant 32 does not affect the spread.

26.5 Linear combinations of random variables

Sometimes the total outcome is best described as a sum of several random variables. For instance, a person's total weekly commute time is the sum of five daily commute times.

A **linear combination** of random variables X and Y is an expression of the form $aX + bY$, where a and b are fixed constants.

John travels to work five days a week. Let X_i represent his travel time on day i. His total weekly commute time is:

$$W = X_1 + X_2 + X_3 + X_4 + X_5$$

If the average daily commute is 18 minutes, the expected weekly total is:

$$E(W) = E(X_1) + E(X_2) + \cdots + E(X_5) = 18 + 18 + 18 + 18 + 18 = 90 \text{ minutes}$$

26.5.1 Expected value of a linear combination

Expected value of a linear combination. For random variables X and Y with constants a and b:

$$E(aX + bY) = a \cdot E(X) + b \cdot E(Y)$$

This generalizes to any number of random variables:

$$E(a_1X_1 + a_2X_2 + \cdots + a_kX_k) = a_1E(X_1) + a_2E(X_2) + \cdots + a_kE(X_k)$$

This property holds whether or not the random variables are independent.

Elena is selling a TV at auction (expected profit X: \$175) and buying a toaster oven (expected cost Y: \$23). Her net gain is $X - Y$. What is her expected net gain?

Show answer

$E(X - Y) = E(X) - E(Y) = 175 - 23 = \$152$.

Leonard has invested \$6,000 in Caterpillar stock (CAT) and \$2,000 in Exxon Mobil stock (XOM). If X represents CAT's monthly percentage return (in decimal form) and Y represents XOM's return, his portfolio change is $6000X + 2000Y$. If $E(X) = 0.020$ and $E(Y) = 0.002$:

$$E(6000X + 2000Y) = 6000 \times 0.020 + 2000 \times 0.002 = \$124$$

Leonard expects to gain \$124 per month on average.

26.5.2 Variance of a linear combination

Variance of a linear combination (independent case). If X and Y are **independent** random variables, then:

$$\text{Var}(aX + bY) = a^2 \cdot \text{Var}(X) + b^2 \cdot \text{Var}(Y)$$

The standard deviation is: $\text{SD}(aX + bY) = \sqrt{a^2\text{Var}(X) + b^2\text{Var}(Y)}$.

Independence is required for the variance formula. The formula $\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$ is only valid when X and Y are independent. If the variables are not independent, an additional covariance term is needed.

However, the expected value formula $E(aX+bY) = aE(X)+bE(Y)$ always holds, regardless of independence.

John's daily commute has a standard deviation of 4 minutes, and the commute times for different days are independent. What is the standard deviation of his weekly commute time $W = X_1 + X_2 + X_3 + X_4 + X_5$?

Each coefficient is 1, and $\text{Var}(X_i) = 4^2 = 16$:

$$\text{Var}(W) = 1^2(16) + 1^2(16) + 1^2(16) + 1^2(16) + 1^2(16) = 5 \times 16 = 80$$

$$\text{SD}(W) = \sqrt{80} \approx 8.94 \text{ minutes}$$

Notice that the standard deviation of the sum (8.94 minutes) is *not* five times the daily standard deviation (which would be 20 minutes). Variabilities do not add up as simply as means do — we must add *variances*, not standard deviations.

Leonard's portfolio has $\text{Var}(X) = 0.0057$ and $\text{Var}(Y) = 0.0021$ (where the stocks are approximately independent). What is the standard deviation of his monthly return?

$$\text{Var}(6000X + 2000Y) = 6000^2 \times 0.0057 + 2000^2 \times 0.0021 = 205{,}200 + 8{,}400 = 213{,}600$$

$$\text{SD} = \sqrt{213{,}600} \approx \$462$$

While an average monthly return of \$124 sounds nice, a standard deviation of \$462 means the returns are highly volatile.

Elena's TV sale has $\text{SD}(X) = \$25$ and her toaster oven purchase has $\text{SD}(Y) = \$8$. Assuming the auctions are independent, compute the standard deviation of her net gain $X - Y$.

Show answer

Note that $X - Y = 1 \cdot X + (-1) \cdot Y$. So $\text{Var}(X - Y) = 1^2 \times 25^2 + (-1)^2 \times 8^2 = 625 + 64 = 689$. Thus $\text{SD}(X - Y) = \sqrt{689} \approx \$26.25$. The negative coefficient is eliminated when we square it — subtracting a random variable introduces *more* variability, not less.

26.5.3 Why variances add even when we subtract

A common source of confusion: the variance of a difference is the *sum* of the variances (not the difference). This makes sense intuitively. If you buy a bag of apples and a bag of oranges, the uncertainty in the total weight is a combination of both uncertainties — regardless of whether you are adding or subtracting the weights. Two *independent* sources of randomness always create more total uncertainty. (If two variables are *positively correlated*, the variance of their difference is smaller than the sum of their variances — which is why paired designs, which compare two measurements on the same subject, can be so precise.)

26.6 Chapter review

26.6.1 Summary

- The **expected value** $E(X) = \sum x_i P(X = x_i)$ is the long-run average of a random variable.
- The **variance** $\text{Var}(X) = \sum (x_i - \mu)^2 P(X = x_i)$ measures spread; the **standard deviation** is $\sigma = \sqrt{\text{Var}(X)}$.
- **Expected value properties**: $E(aX + b) = aE(X) + b$ (always holds).
- **Variance properties**: $\text{Var}(aX + b) = a^2\text{Var}(X)$ (adding a constant does not change variance).
- **Linear combinations**: $E(aX+bY) = aE(X)+bE(Y)$ (always). $\text{Var}(aX+bY) = a^2\text{Var}(X)+b^2\text{Var}(Y)$ (requires independence).
- For independent random variables, variances always add, even when the random variables are subtracted.

26.7 Exercises

1. **Portfolio return.** A portfolio's value increases by 18% during a financial boom and by 9% during normal times. It decreases by 12% during a recession. What is the expected return on this portfolio if each scenario is equally likely?

2. **Baggage fees.** An airline charges the following baggage fees: \$25 for the first bag and \$35 for the second. Suppose 54% of passengers have no checked luggage, 34% have one piece of checked luggage and 12% have two pieces. We suppose a negligible portion of people check more than two bags.

a. Build a probability model, compute the average revenue per passenger, and compute the corresponding standard deviation.

b. About how much revenue should the airline expect for a flight of 120 passengers? With what standard deviation? Note any assumptions you make and if you think they are justified.

3. **American roulette.** The game of American roulette involves spinning a wheel with 38 slots: 18 red, 18 black, and 2 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball. Gamblers can place bets on red or black. If the ball lands on their color, they double their money. If it lands on another color, they lose their money. Suppose you bet \$1 on red. What's the expected value and standard deviation of your winnings?

4. **One big bet vs many small bets.** A college basketball player makes 55% of her free throws. Consider a game where you bet on her next attempt: if she makes the shot you win the amount you bet; if she misses you lose that amount.

a. **Setup A.** You place a single \$3 bet on one free throw. Find the expected value and standard deviation of your winnings W_A.

b. **Setup B.** You place three independent \$1 bets, one on each of three free-throw attempts. Find the expected value and standard deviation of your total winnings W_B.

c. Compare your answers to parts (a) and (b). What does the comparison say about the riskiness of concentrating a stake in one large bet versus splitting it across several small ones?

5. **Cost of breakfast.** Sally gets a cup of coffee and a muffin every day for breakfast from one of the many coffee shops in her neighborhood. She picks a coffee shop each morning at random and independently of previous days. The average price of a cup of coffee is \$1.40 with a standard deviation of 30¢ (\$0.30), the average price of a muffin is \$2.50 with a standard deviation of 15¢, and the two prices are independent of each other.

a. What is the mean and standard deviation of the amount she spends on breakfast daily?

b. What is the mean and standard deviation of the amount she spends on breakfast weekly (7 days)?

6. **Scooping ice cream.** Ice cream usually comes in 1.5 quart boxes (48 fluid ounces), and ice cream scoops hold about 2 ounces. However, there is some variability in the amount of ice cream in a box as well as the amount of ice cream scooped out. We represent the amount of ice cream in the box as X and the amount scooped out as Y. Suppose these random variables have the following means, standard deviations, and variances:

Y | 2 | 0.25 | 0.0625 |

a. An entire box of ice cream, plus 3 scoops from a second box is served at a party. How much ice cream do you expect to have been served at this party? What is the standard deviation of the amount of ice cream served?

b. How much ice cream would you expect to be left in the box after scooping out one scoop of ice cream? That is, find the expected value of $X - Y$. What is the standard deviation of the amount left in the box?

c. Using the context of this exercise, explain why we add variances when we subtract one random variable from another.

7. **Variance of a mean, Part I.** Suppose X_1 and X_2 are independent observations from a population with mean μ and variance $\sigma^2 = 25$. Let $\bar{X} = (X_1 + X_2)/2$.

a. Compute Var$(X_1 + X_2)$ and Var$(\bar{X})$.

b. In general (for any σ^2), which formula gives Var$(\bar{X})$ when $n = 2$?

8. **Variance of a mean, Part II.** Now suppose X_1, X_2, X_3 are independent observations from a population with variance $\sigma^2 = 25$. Let $\bar{X} = (X_1 + X_2 + X_3)/3$.

a. Compute Var$(\bar{X})$.

b. Compared to Part I (where $n = 2$), as the sample size n increases, what happens to Var$(\bar{X})$?

c. For a sample of size n from this same population, what is the smallest n that guarantees SD$(\bar{X}) \leq 1$?

9. **Variance of a mean, Part III.** Generalize Parts I and II. Suppose $X_1, X_2, \ldots, X_n$ are independent observations from a population with mean μ and variance σ^2, and let $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$.

a. Which formula gives Var$(\bar{X})$?

b. Suppose $\sigma^2 = 36$ and $n = 16$. Compute Var$(\bar{X})$ and SD$(\bar{X})$.

Chapter 27

Binomial Distribution

How many of 400 randomly sampled adults are left-handed? How many of 20 randomly selected parts from an assembly line are defective? These "count of successes" questions arise constantly in statistics and are answered by the **binomial distribution**. In this chapter, we identify the four conditions that define a binomial setting, develop the formula for computing binomial probabilities, derive the mean and standard deviation, and show how the normal distribution can approximate binomial probabilities when the sample size is large.

27.1 The binomial setting

The binomial distribution describes the number of successes in a fixed number of independent trials, where each trial has only two possible outcomes.

The four binomial conditions. A random variable X follows a binomial distribution if:

1. **Fixed number of trials**: The number of trials n is fixed in advance.
2. **Binary outcomes**: Each trial results in one of two outcomes, which we label "success" and "failure."
3. **Independence**: The trials are independent — the outcome of one trial does not affect the outcome of any other trial.
4. **Constant probability**: The probability of success p is the same for every trial.

If all four conditions are met, we say X follows a binomial distribution with parameters n and p, written $X \sim \text{Binomial}(n, p)$.

A health insurance company found that 70% of the people they insure stay below their deductible in any given year. Consider a random sample of four individuals. Does the number who stay below their deductible follow a binomial distribution?

1. **Fixed** n: Yes, $n = 4$ individuals.
2. **Binary**: Yes, each person either stays below the deductible (success) or exceeds it (failure).
3. **Independence**: Yes, because the sample is random from a large population.
4. **Constant** p: Yes, each person has a 70% chance of staying below the deductible.

All four conditions are met, so $X \sim \text{Binomial}(n = 4, p = 0.70)$.

The labels "success" and "failure" are just conventions — a "success" does not have to be a good outcome. If we are counting defective parts, a "success" might be finding a defective part. The mathematical framework does not depend on which outcome we call a success, as long as we are consistent.

The probability that a random smoker will develop a severe lung condition is about 0.3. If you have 4 friends who smoke, are the conditions for the binomial model satisfied?

Show answer

This is questionable. If the friends know each other, the independence condition may be violated. Acquaintances may have similar smoking habits or might influence each other. If, however, we could treat them as a random sample of smokers, the conditions would be approximately met.

27.1.1 The Bernoulli distribution: a single trial

A single trial with two outcomes is called a **Bernoulli trial**, and the resulting random variable is a **Bernoulli random variable**.

If X is a random variable that takes value 1 with probability p (success) and 0 with probability $1 - p$ (failure), then X is a **Bernoulli random variable** with:

$$\mu = p \qquad \sigma = \sqrt{p(1-p)}$$

A binomial random variable is simply the sum of n independent Bernoulli random variables, each with the same probability p.

27.2 Computing binomial probabilities

27.2.1 Building the formula

Let's find the probability that exactly 1 of 4 randomly selected insured individuals exceeds the deductible (so 3 "successes" — staying below — and 1 "failure").

Consider the specific scenario where person A exceeds and persons B, C, D do not:

$$P(A = \text{exceed}, B = \text{not}, C = \text{not}, D = \text{not}) = (0.3)(0.7)(0.7)(0.7) = (0.7)^3(0.3)^1 = 0.103$$

But there are four such scenarios — any one of the four people could be the one who exceeds. Each scenario has the same probability. So:

$$P(\text{exactly 3 successes in 4 trials}) = 4 \times (0.7)^3(0.3)^1 = 0.412$$

The general pattern is:

$$P(X = k) = [\text{number of arrangements}] \times [\text{probability of one arrangement}]$$

27.2.2 The binomial coefficient

The number of ways to arrange k successes in n trials is given by the **binomial coefficient**:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

read as "n choose k," where $n!$ (read "n factorial") means $n \times (n-1) \times \cdots \times 2 \times 1$, with the special case $0! = 1$.

For our example: $\binom{4}{3} = \frac{4!}{3!\cdot 1!} = \frac{24}{6\times 1} = 4$, confirming the four arrangements we identified.

27.2.3 The binomial probability formula

Binomial probability formula. If $X \sim \text{Binomial}(n, p)$, the probability of exactly k successes is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!}\, p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots, n$.

What is the probability that 5 of 8 randomly selected insured individuals will not exceed the deductible?

With $n = 8$, $k = 5$, and $p = 0.70$:

$$P(X = 5) = \binom{8}{5}(0.70)^5(0.30)^3$$

Computing the binomial coefficient:

$$\binom{8}{5} = \frac{8!}{5! \cdot 3!} = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = 56$$

And the probability of a single arrangement: $(0.70)^5(0.30)^3 \approx 0.00454$.

$$P(X = 5) = 56 \times 0.00454 \approx 0.254$$

Suppose 4 friends who smoke can be treated as a random sample, with $P(\text{severe lung condition}) = 0.3$ for each. (a) What is the probability that none develop a severe lung condition? (b) What is the probability that exactly one does? (c) What is the probability that at most one does?

Show answer

(a) $P(X = 0) = \binom{4}{0}(0.3)^0(0.7)^4 = 1 \times 1 \times 0.2401 = 0.2401$. (b) $P(X = 1) = \binom{4}{1}(0.3)^1(0.7)^3 = 4 \times 0.3 \times 0.343 = 0.4116$. (c) $P(X \leq 1) = P(X = 0) + P(X = 1) = 0.2401 + 0.4116 = 0.6517$.

What is the probability that at least 2 of the 4 smoking friends will develop a severe lung condition?

Show answer

Using the complement: $P(X \geq 2) = 1 - P(X \leq 1) = 1 - 0.6517 = 0.3483$.

27.2.4 Computing binomial probabilities in practice

For small values of n and k, the binomial formula can be computed by hand. For larger problems, use technology:

- **StatLens**: The Binomial Distribution calculator computes probabilities and visualizes the distribution.
- **Calculator**: Most graphing calculators have a `binompdf` and `binomcdf` function.
- **Software**: Any statistical software can compute binomial probabilities. Look for functions named something like "binomial CDF" or "binomial probability."

For example, to find $P(X \leq 42)$ when $X \sim \text{Binomial}(400, 0.15)$, computing 43 individual probabilities by hand would be impractical. Using the StatLens Binomial calculator with $n = 400$, $p = 0.15$, and $P(X \leq 42)$, we get approximately 0.0054.

27.3 Mean and standard deviation

Mean and standard deviation of a binomial distribution. If $X \sim \text{Binomial}(n, p)$, then:

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

These formulas follow from the fact that a binomial random variable is the sum of n independent Bernoulli random variables, each with mean p and variance $p(1-p)$.

If 40 individuals are randomly sampled and each has a 70% chance of staying below their insurance deductible, how many would you expect to stay below? What is the standard deviation?

$$\mu = np = 40 \times 0.70 = 28$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{40 \times 0.70 \times 0.30} = \sqrt{8.4} \approx 2.9$$

We expect about 28 individuals to stay below their deductible, give or take about 2.9. Using the rough "95% within 2 standard deviations" guideline, we would expect between about 22 and 34 individuals in most samples.

Suppose 7 friends who smoke can be treated as a random sample, with $p = 0.3$ for developing a severe lung condition. How many would you expect to develop the condition? What is the standard deviation?

Show answer

$\mu = np = 7 \times 0.3 = 2.1$. $\sigma = \sqrt{7 \times 0.3 \times 0.7} = \sqrt{1.47} \approx 1.21$. You would expect about 2 of the 7 friends to develop a severe lung condition.

27.3.1 Simulation perspective: Does the formula match reality?

We can verify the binomial formulas using simulation. Consider $X \sim \text{Binomial}(n = 20, p = 0.3)$:

- **Theory**: $\mu = 20 \times 0.3 = 6$ and $\sigma = \sqrt{20 \times 0.3 \times 0.7} = \sqrt{4.2} \approx 2.05$
- **Simulation**: Generate 10,000 binomial random samples, each with $n = 20$ and $p = 0.3$. The sample mean of the 10,000 counts will be very close to 6, and the sample standard deviation will be very close to 2.05.

This is another instance of the Law of Large Numbers: the simulated distribution converges to the theoretical distribution as we increase the number of simulations.

See it in action. Open the Sampling Distribution Lab and draw many samples of size 20. Each sample's $\hat{p} = X/20$ lands in the distribution, so the histogram is a simulated binomial distribution (rescaled by $1/20$). Compare it with the exact PMF in the Binomial Distribution Explorer.

27.4 Normal approximation to the binomial

The binomial formula can be cumbersome for large n, especially when computing cumulative probabilities like $P(X \leq 42)$ for $n = 400$. Fortunately, when n is large enough, the binomial distribution is well approximated by the normal distribution.

Approximately 15% of the US population smokes cigarettes. A local government surveyed 400 randomly selected individuals and found only 42 smokers. If the true proportion is 0.15, what is the probability of observing 42 or fewer smokers?

The exact answer requires summing 43 binomial probabilities, which gives $P(X \leq 42) = 0.0054$.

As the sample size increases, the binomial distribution becomes increasingly bell-shaped and symmetric. This is illustrated in the figure below.

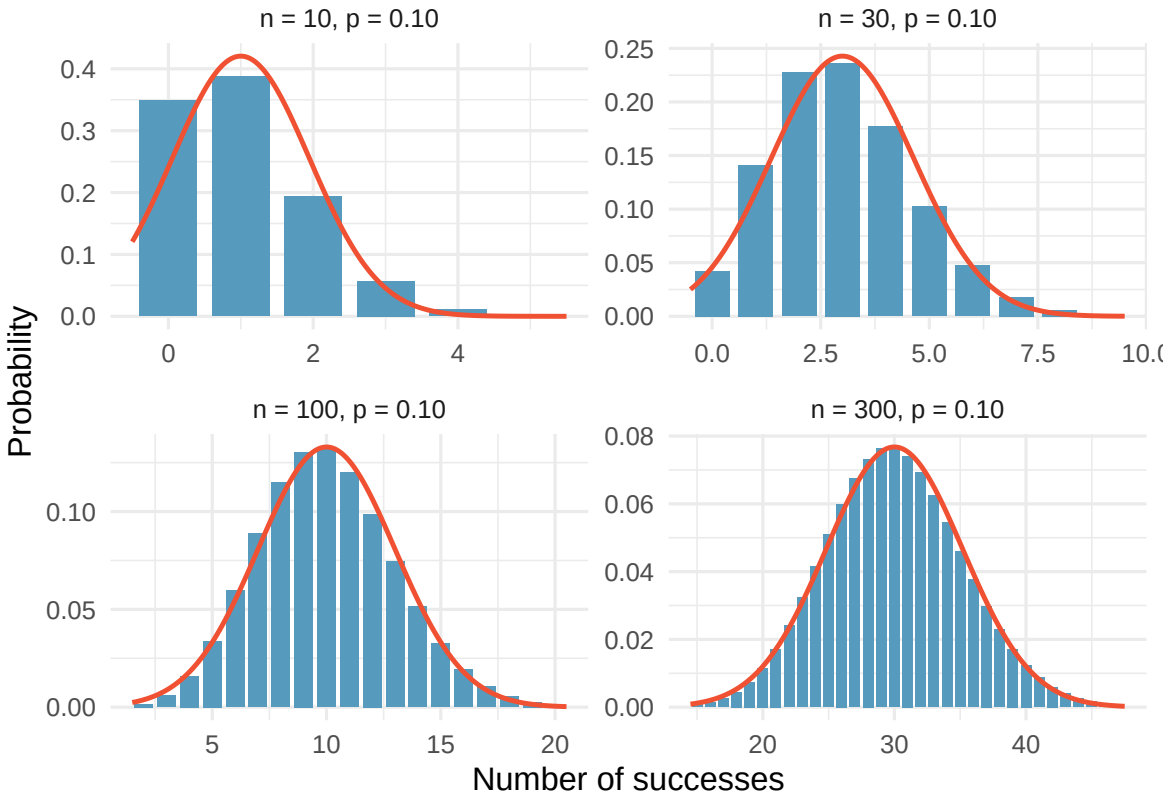


Figure 27.1: Binomial distributions with p = 0.10 and sample sizes n = 10, 30, 100, and 300. As n increases, the distribution approaches a normal curve. ↗ Try this live — drive n and p in the Binomial PMF and watch the shape settle.

Normal approximation to the binomial. The binomial distribution with parameters n and p is approximately normal when both $np \geq 10$ and $n(1-p) \geq 10$. The approximating normal distribution has:

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

Use the normal approximation to estimate $P(X \leq 42)$ for $X \sim \text{Binomial}(400, 0.15)$.

First, check the conditions: $np = 400 \times 0.15 = 60 \geq 10$ and $n(1-p) = 400 \times 0.85 = 340 \geq 10$. The normal approximation is appropriate.

The approximating distribution is $N(\mu = 60, \sigma = 7.14)$, where $\sigma = \sqrt{400 \times 0.15 \times 0.85} = 7.14$.

Compute the Z-score:

$$Z = \frac{42 - 60}{7.14} = \frac{-18}{7.14} = -2.52$$

Using the standard normal distribution: $P(Z \leq -2.52) = 0.0059$.

This is very close to the exact binomial probability of 0.0054.

A coin is flipped 100 times. Use the normal approximation to find the probability of getting 60 or more heads.

Show answer

Check conditions: $np = 100 \times 0.5 = 50 \geq 10$ and $n(1-p) = 50 \geq 10$. The approximation is appropriate. $\mu = 50$, $\sigma = \sqrt{100 \times 0.5 \times 0.5} = 5$. $Z = (60 - 50)/5 = 2$. $P(Z \geq 2) = 1 - 0.9772 = 0.0228$. There is about a 2.3% chance of getting 60 or more heads in 100 flips of a fair coin.

27.4.1 Connecting to inference

The normal approximation to the binomial is not just a computational shortcut — it is the theoretical basis for the inference methods we developed in Parts II and III. When we studied the sampling distribution of the sample proportion $\hat{p}$ in the sampling distributions chapter, we were implicitly using the fact that the count of successes X in a sample of size n follows a binomial distribution, and that this binomial distribution is approximately normal for large n.

Specifically, if $X \sim \text{Binomial}(n, p)$, then $\hat{p} = X/n$, and the normal approximation tells us:

$$\hat{p} \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

This is exactly the result we used when constructing confidence intervals and performing hypothesis tests for proportions in the inference for proportions chapter. The theory developed in this chapter *explains* why those methods work.

27.5 Chapter review

27.5.1 Summary

- The **binomial distribution** counts the number of successes in n independent trials, each with the same probability of success p.
- **Four conditions**: fixed n, binary outcomes, independence, constant p.
- **Binomial probability formula**: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- **Mean and standard deviation**: $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.
- When $np \geq 10$ and $n(1-p) \geq 10$, the binomial distribution is well approximated by the normal distribution $N(np, \sqrt{np(1-p)})$.
- The normal approximation to the binomial provides the theoretical justification for the inference methods on proportions from the inference for proportions chapter.

27.6 Exercises

1. **Underage drinking, Part I.** Data collected by the Substance Abuse and Mental Health Services Administration (SAMSHA) suggests that 69.7% of 18-20 year olds consumed alcoholic beverages in any given year.

a. Suppose a random sample of ten 18-20 year olds is taken. Is the use of the binomial distribution appropriate for calculating the probability that exactly six consumed alcoholic beverages? Explain.

b. Calculate the probability that exactly 6 out of 10 randomly sampled 18- 20 year olds consumed an alcoholic drink.

c. What is the probability that exactly four out of ten 18-20 year olds have *not* consumed an alcoholic beverage?

d. What is the probability that at most 2 out of 5 randomly sampled 18-20 year olds have consumed alcoholic beverages?

e. What is the probability that at least 1 out of 5 randomly sampled 18-20 year olds have consumed alcoholic beverages?

2. **Chickenpox, Part I.** Boston Children's Hospital estimates that 90% of Americans have had chickenpox by the time they reach adulthood.

a. Suppose we take a random sample of 100 American adults. Is the use of the binomial distribution appropriate for calculating the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood? Explain.

b. Calculate the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood.

c. What is the probability that exactly 3 out of a new sample of 100 American adults have *not* had chickenpox in their childhood?

d. What is the probability that at least 1 out of 10 randomly sampled American adults have had chickenpox?

e. What is the probability that at most 3 out of 10 randomly sampled American adults have *not* had chickenpox?

3. **Underage drinking, Part II.** We learned in a previous exercise that about 70% of 18-20 year olds consumed alcoholic beverages in any given year. We now consider a random sample of fifty 18-20 year olds.

a. How many people would you expect to have consumed alcoholic beverages? And with what standard deviation?

b. Would you be surprised if there were 45 or more people who have consumed alcoholic beverages?

c. What is the probability that 45 or more people in this sample have consumed alcoholic beverages? How does this probability relate to your answer to part (b)?

4. **Chickenpox, Part II.** We learned in a previous exercise that about 90% of American adults had chickenpox before adulthood. We now consider a random sample of 120 American adults.

a. How many people in this sample would you expect to have had chickenpox in their childhood? And with what standard deviation?

b. Would you be surprised if there were 105 people who have had chickenpox in their childhood?

c. What is the probability that 105 or fewer people in this sample have had chickenpox in their childhood? How does this probability relate to your answer to part (b)?

5. **Four-color spinner.** A carnival booth has a four-color spinner divided into four equal sectors: red, blue, green, and yellow. You spin the wheel three times. Each spin is independent and each color has probability $1/4$ per spin. Calculate the probability of getting

a. at least one red in 3 spins.

b. exactly 2 reds.

c. exactly 1 blue.

d. at most 2 greens.

6. **Arachnophobia.** A Gallup Poll found that 7% of teenagers (ages 13 to 17) suffer from arachnophobia and are extremely afraid of spiders. At a summer camp there are 10 teenagers sleeping in each tent. Assume that these 10 teenagers are independent of each other.%

a. Calculate the probability that at least one of them suffers from arachnophobia.

b. Calculate the probability that exactly 2 of them suffer from arachnophobia.

c. Calculate the probability that at most 1 of them suffers from arachnophobia.

d. If the camp counselor wants to make sure no more than 1 teenager in each tent is afraid of spiders, does it seem reasonable for him to randomly assign teenagers to tents?

7. **Eye color, Part II.** a previous exercise introduces a husband and wife with brown eyes who have 0.75 probability of having children with brown eyes, 0.125 probability of having children with blue eyes, and 0.125 probability of having children with green eyes.

a. What is the probability that their first child will have green eyes and the second will not?

b. What is the probability that exactly one of their two children will have green eyes?

c. If they have six children, what is the probability that exactly two will have green eyes?

d. If they have six children, what is the probability that at least one will have green eyes?

e. What is the probability that the first green eyed child will be the 4th child?

f. Would it be considered unusual if only 2 out of their 6 children had brown eyes?

8. **Sickle cell anemia.** Sickle cell anemia is a genetic blood disorder where red blood cells lose their flexibility and assume an abnormal, rigid, "sickle" shape, which results in a risk of various complications. If both parents are carriers of the disease, then a child has a 25% chance of having the disease, 50% chance of being a carrier, and 25% chance of neither having the disease nor being a carrier. If two parents who are carriers of the disease have 3 children, what is the probability that

a. two will have the disease?

b. none will have the disease?

c. at least one will neither have the disease nor be a carrier?

d. the first child with the disease will the be 3rd child?

9. **Playlist orderings.** A music app has 8 different songs in a playlist and plays them in a random order with no repeats — every distinct ordering is equally likely.

a. How many distinct orderings of the 8 songs are possible?

b. What is the probability that "Track 1" is played first?

c. What is the probability that "Track 1" is played first AND "Track 2" is played last?

10. **Allergic campers in a cabin.** At a summer camp, cabins hold 4 campers. Independently for each camper, the probability of being allergic to a particular plant is 0.15.

a. The orderings of (A = allergic, N = not allergic) for a cabin of 4 that have exactly 2 allergic campers are: AANN, ANAN, ANNA, NAAN, NANA, NNAA. How many distinct orderings is that?

b. Compute the probability of one specific ordering with exactly 2 allergic campers (for example, AANN).

c. Use the binomial formula to compute the probability that exactly 2 of the 4 campers are allergic.

d. Multiplying your answers from (a) and (b) should give the answer to (c). Explain why this works.

11. **Airline overbooking.** A regional airline runs flights on a small commuter plane with 50 seats. Historically, 7% of booked passengers do not show up, so the airline sells 54 tickets for each flight. Let X be the number of ticketed passengers who actually show up, modeled as $X \sim \text{Binomial}(n = 54, p = 0.93)$.

a. Find the mean μ of X.

b. Find the standard deviation σ of X.

c. Using the normal approximation with continuity correction, estimate the probability the flight is overbooked, $P(X > 50)$.

d. Interpret your answer to part (c) in plain language.

12. **Email campaign opens.** A marketing team sends a promotional email to 8,000 independent recipients. Historically, each recipient opens the email with probability 0.12, and opens are independent. Let X be the number of recipients who open the email, so $X \sim \text{Binomial}(n = 8000, p = 0.12)$. The team wants the probability that at least 1,000 recipients open the email.

a. Compute $\mu = E(X)$ and $\sigma = \text{SD}(X)$.

b. Using the normal approximation with continuity correction, find the z-score corresponding to "at least 1,000 opens."

c. Approximate $P(X \geq 1000)$.

d. State the condition that makes the normal approximation appropriate here.

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Does this scenario fit the binomial?** The binomial requires (i) a fixed number of trials n, (ii) two outcomes per trial, (iii) constant success probability p, and (iv) independence across trials. For each scenario, check the four conditions and decide whether the binomial applies.
   a. Flip a fair coin 50 times; count the number of heads.
   b. Draw 5 cards *without replacement* from a deck; count the number of aces.
   c. Survey 100 randomly selected adults; count how many own a smartphone.
   d. Roll a die until you get the first 6; count the number of rolls.
   e. Test 200 light bulbs from one production line; count how many fail within 100 hours.
2. **Read the binomial PMF.** Open the Binomial tool. Set $n = 20$ and $p = 0.3$.
   a. Report $P(X = 6)$ (the bar at $X = 6$). Where is the mode of the distribution?
   b. Compute the mean and standard deviation from the formulas $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. How close is the visible "center" of the bars to your computed μ?
   c. Switch to **cumulative** mode. What is $P(X \leq 5)$? What is $P(X \geq 8)$?
   d. Toggle on the **normal approximation overlay**. Does it match the binomial bars well at $n = 20$, $p = 0.3$? Why or why not?
3. **The normal approximation to the binomial — when does it work?** Still in the Binomial tool.
   a. Set $n = 20$, $p = 0.05$ (rare event). What does the PMF look like — symmetric or skewed? Turn on the normal overlay. Does the normal model fit?
   b. Now $n = 200$, $p = 0.05$. PMF shape? Normal fit?
   c. State the **success–failure rule**: the normal approximation works well when $np \geq 10$ and $n(1-p) \geq 10$. Check this for cases (a) and (b).
   d. Why does the rule use *both* $np \geq 10$ **and** $n(1-p) \geq 10$ rather than just one? (Hint: consider $p = 0.99$, $n = 100$.)
4. **Connecting the binomial to one-proportion inference.** A study of 200 randomly chosen tablet users finds 14 who experienced a side effect, $\hat{p} = 14/200 = 0.07$.
   a. Under $H_0 : p = 0.10$, the count X of side effects is $X \sim \text{Binomial}(200, 0.10)$. Use the Binomial tool in **cumulative** mode to find the exact $P(X \leq 14)$.
   b. Now use the One-Proportion z-Test tool with this scenario, $H_A : p < 0.10$. Report the (normal-approximation) p-value. How close is it to the binomial's exact answer in (a)?
   c. Check the success–failure condition under H_0. Should you trust the z-test's p-value?
   d. In one sentence, when would you prefer the **exact** binomial probability over the **approximate** z-test p-value?
5. **Expectation and the law of large numbers.** A casino offers a game with $n = 100$ trials, win probability $p = 0.45$ per trial, and a \$1 wager per trial.
   a. Open Binomial $n = 100, p = 0.45$. What's the **expected number of wins**? What's the probability you come out **ahead** (more than 50 wins)?
   b. Now consider playing 100 *rounds* of this game (100 sessions of 100 trials each). What does the **law of large numbers** predict about your average per-round winnings as the number of rounds grows?
   c. A player says: "In each round my expected loss is small ($-\$10$), so I should keep playing." Why is this the **gambler's mathematical mistake** (not just a sentiment)?
   d. How does the binomial probability of "coming out ahead" change as n grows from 100 to 1000 to 10000 trials? Why?

Chapter 28

More About the Normal Distribution

Builds on the normal-approximation chapter. This chapter assumes you've already met the normal distribution as it was developed in the normal-approximation chapter — the $N(\mu, \sigma)$ notation, the 68-95-99.7 rule, z-scores, and finding tail probabilities and percentiles. Here we look at two extensions that didn't fit earlier: **how to check whether data actually are approximately normal**, and **which real-world variables the normal model fits well (and which it doesn't).**

The normal distribution is powerful precisely because so many real-world variables are approximately normal. But *approximately* is a word doing a lot of work — some variables really are close to normal, others are clearly not, and it matters which is which before you use the normal model to compute probabilities or run a formula-based test. This chapter gives you two tools for making that call.

28.1 Checking whether data are normal

Before using the normal distribution as a model, we should check whether the assumption of normality is reasonable. Two visual methods are commonly used, and they complement each other.

28.1.1 Method 1: Histogram with a normal curve overlay

The simplest approach is to plot a histogram of the data and overlay the best-fitting normal curve (using $\bar{x}$ for μ and s for σ). If the histogram closely matches the curve, the normal model is reasonable. If the histogram is skewed, bimodal, or has long tails the curve doesn't capture, the model is a poor fit.

This works well when n is large enough for the histogram to reveal the true shape — typically $n \geq 50$ or so. With smaller samples, the histogram is noisy and even genuinely normal data can look ragged.

28.1.2 Method 2: Normal probability plots (QQ plots)

A **normal probability plot** (also called a **QQ plot**, short for "quantile-quantile plot") plots the observed data values against the values we would expect if the data were perfectly normal. If the data are approximately normal, the points will fall close to a straight line.

How to read a QQ plot:

- **Points on or near the line**: The data are consistent with a normal distribution.
- **Points curving up at the right end**: The distribution is **right-skewed** (a long upper tail).
- **Points curving down at the left end**: The distribution is **left-skewed** (a long lower tail).
- **Points curving away from the line at both ends**: The distribution has **heavier tails** than normal — more extreme values than the normal predicts.

Figure 28.1 shows the four patterns you will see most often.

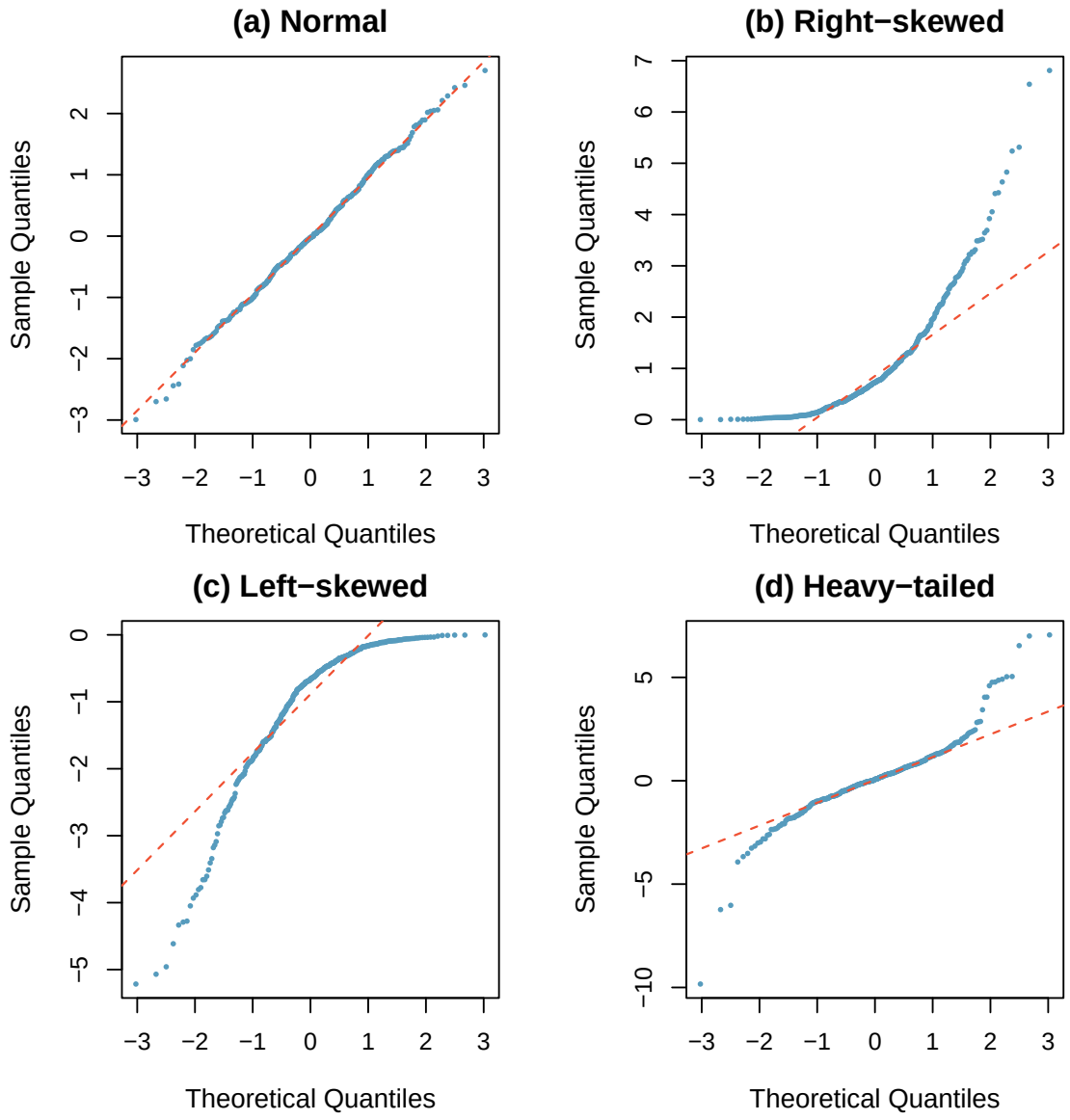


Figure 28.1: Four QQ plots illustrating common patterns: (a) approximately normal data — points fall close to the line; (b) right-skewed data — points curve upward at the right; (c) left-skewed data — points curve downward at the left; (d) heavy-tailed data — points curve away from the line at both ends.

Sample size matters. With small sample sizes ($n < 30$), we should expect more variability in the QQ plot even when the data truly come from a normal distribution. Deviations must be more severe to be considered evidence against normality when n is small.

A researcher collects data and creates a QQ plot. The points follow the diagonal line closely in the middle of the distribution but curve upward at the right end. What does this suggest about the shape of the data?

Show answer

The QQ plot suggests the data are right-skewed. The bulk of the data are approximately normal, but there are larger values in the upper tail than the normal distribution would predict.

28.2 When to use — and not use — the normal model

The normal distribution is a good model for data that are:

- **Symmetric**: The histogram is roughly mirror-image around the center.
- **Unimodal**: There is a single peak.
- **Bell-shaped**: The histogram has the characteristic tapered shape of the normal curve.
- **No extreme outliers**: Outliers distort both the mean and standard deviation, making the normal model a poor fit.

Many variables are approximately normal:

- Standardized test scores (SAT, ACT, GRE)
- Heights of adults (within a sex)
- Measurement errors in scientific instruments
- Many biological measurements (blood pressure, cholesterol levels)
- Averages of many small independent effects — which is why sample means behave normally by the Central Limit Theorem

Many variables are **not** normal:

- **Income distributions** — strongly right-skewed; a few very high earners pull the tail out.
- **Waiting times** — typically right-skewed, often bounded below by zero.
- **Count data** — usually better modeled by the binomial or Poisson, not the normal.
- **Individual values that are percentages** (e.g., each person's body-fat percentage) — bounded, so they can't be strictly normal. (A sample proportion $\hat{p}$ from a large sample is different: as a *statistic*, it is approximately normal.)
- **Ages** — bounded below by 0 and often bounded above by natural limits, so the tails don't behave like a normal.

When the normal model doesn't fit, don't force it. For skewed data, consider a transformation (e.g., log for right-skewed positive data), a different model (binomial for counts of successes, Poisson for rare event counts), or a **simulation-based method** like the bootstrap or a randomization test. Those simulation methods (developed in the randomization-tests and bootstrap-CI chapters) don't require normality — though with very small samples from skewed populations they also tend to give intervals that are too narrow.

A hospital measures the time patients wait to be seen in the emergency room. The data range from 3 minutes to 6 hours, with most patients waiting under 45 minutes but a long right tail. A researcher wants to run a one-sample t-test to see if the mean wait time exceeds 30 minutes. Is a t-test appropriate?

The data are strongly right-skewed and bounded below by zero, so they are clearly not normal. A one-sample t-test *may* still be reasonable if the sample size is large — the CLT says the sampling distribution of $\bar{x}$ will be approximately normal even when the population is skewed. But with a small sample (say, $n < 30$), the t-test could give misleading results. A bootstrap interval avoids assuming a normal population but does not rescue a tiny sample: with $n = 8$ from a strongly right-skewed population, StatLens's CI Coverage Simulator shows 95% percentile intervals capturing the mean only about 84% of the time (about 88% for the t-interval). With strongly skewed data, the real remedy is more data.

See it in action. Open the Descriptive Statistics Explorer on any bundled dataset — switch to the histogram view to eyeball symmetry and unimodality, then check the sample statistics for extreme outliers. For a QQ plot, open the One-Sample t-Test page, load a dataset, and switch its conditions panel to **QQ Plot**.

28.3 Chapter review

28.3.1 Summary

- Before applying the normal model, **check whether the data are approximately normal**. Two visual tools help: a **histogram with a normal-curve overlay**, and a **QQ plot** (points near the diagonal = approximately normal; systematic curvature = departure from normality).
- The four QQ-plot patterns worth recognizing: **normal** (straight), **right-skewed** (curves up at the right), **left-skewed** (curves down at the left), and **heavy-tailed** (curves away at both ends).
- The normal model fits **symmetric, unimodal, bell-shaped** variables without extreme outliers. Standardized test scores, adult heights, and measurement errors are canonical examples.
- The normal model does **not** fit strongly skewed data (income, waiting times), bounded data (proportions, ages), or count data (better handled by the binomial or Poisson).
- When the normal model doesn't fit, use a **simulation-based** method (bootstrap CI, randomization test) or a **different distribution family** appropriate for the data — and remember that no method rescues a very small sample from a strongly skewed population.

28.3.2 Key terms

QQ plot (normal probability plot), normality assessment, histogram-with-normal-overlay, symmetric / unimodal / bell-shaped, heavy-tailed distribution, right-skewed / left-skewed, bootstrap alternative to normal-based inference.

28.4 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Reading a QQ plot.** For each described pattern below, name the shape of the underlying data distribution and briefly explain how you can tell.

a. Points fall very close to the reference line across the whole range.

b. Points hug the line in the middle but curve **upward** away from the line at the right end.

c. Points hug the line in the middle but curve **downward** away from the line at the left end.

d. Points fall below the line at the left end **and** above the line at the right end.

2. **Which of these variables would you model as normal?** For each variable, decide whether the normal distribution is a reasonable model for individual observations. Answer *yes*, *no*, or *depends*, and explain in one sentence.

a. Adult male heights in a large city.

b. Household incomes in the United States.

c. The number of car accidents at a given intersection per day.

d. The average of many random draws from any distribution.

e. The proportion of a small sample that responds "yes" to a survey question.

f. Systolic blood pressure of healthy adults.

3. **Small-sample QQ plot interpretation.** A researcher has a sample of $n = 15$ measurements. The QQ plot shows some scatter around the reference line — the points don't fall perfectly on it, but there's no clear systematic curve either.

a. Is this stronger evidence *for* or *against* the data being approximately normal? Explain.

b. If the same amount of scatter appeared in a QQ plot with $n = 500$, would your conclusion change? Why?

c. A colleague says "the points don't fall exactly on the line, so we can't use a t-test." Is this reasoning sound? Explain.

4. **Pick a method.** For each scenario, decide whether a **one-sample t-test** is reasonable, or whether a **bootstrap CI** (or different approach) would be safer. Explain in one sentence.

a. You want to estimate the mean commute time in a city. A histogram of your $n = 40$ commute-time observations is roughly symmetric with no outliers.

b. You want to estimate the mean insurance claim amount. Your $n = 25$ observations are strongly right-skewed with three extreme outliers.

c. You want to estimate the mean systolic blood pressure of adult women. Your $n = 200$ observations look approximately normal.

d. You want to test whether the mean wait time at an emergency room exceeds 30 minutes. Your $n = 8$ observations are strongly right-skewed and bounded below by 0.

5. **Two normality checks, one dataset.** You have a dataset of daily rainfall (inches) for 300 days at a weather station. You plot both a histogram with a fitted normal overlay and a QQ plot. The histogram is strongly right-skewed with many days near 0 and a few days with heavy rain; the QQ plot's points curve sharply upward at the right end.

a. Do the two plots agree or disagree on normality?

b. Suggest one transformation of the data that might make the transformed values more nearly normal, and explain why it would help.

c. Suppose you compute a 95% one-sample t confidence interval for the mean daily rainfall from these 300 observations. Given the CLT and $n = 300$, do you trust it? What's the caveat?

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **The empirical rule, checked against the actual normal.** Open the Normal tool preset to $\mu = 0, \sigma = 1$.
   a. Set the middle-region boundaries to ± 1. What is the area? Is it 68%?
   b. Now ± 2. Is it 95%?
   c. Now ± 3. Is it 99.7%?
   d. The empirical rule says "68–95–99.7." Where do the exact percentages come from — the rule or the curve?

2. **Standardizing across different normals.** Adult resting heart rates are approximately $N(\mu = 72, \sigma = 12)$ beats per minute. Adult body temperatures are approximately $N(\mu = 98.6, \sigma = 0.7)$ °F.
   a. A patient has a heart rate of 90. What is the patient's z-score?
   b. Another patient has a body temperature of 99.6 °F. What is that patient's z-score?
   c. Which patient is *more* unusual relative to their respective population? Why does the z-score answer this even though the two measurements are in different units?
   d. Use the standard normal z tool to find the percentile rank of each patient (the area to the left of their z-score). Are they in roughly the same percentile?

3. **Inverse-normal (working backward).** A standardized test has scores $\sim N(500, 100)$.
   a. What score corresponds to the **top 10%** of test-takers? Use the normal tool to set the right-tail area to 0.10 and read off the boundary.
   b. What score is at the **25th percentile**? Use the left-tail area of 0.25.
   c. What is the **interquartile range** of the score distribution?
   d. A scholarship is offered to students scoring above the **90th percentile**. The college says: "Half of our applicants miss out by 50 points or fewer." Is this surprising, given the distribution?

4. **Within vs. between normals — a common confusion.** A factory measures the diameters of bolts. The population of individual diameters is $N(\mu = 10.0\,\text{mm}, \sigma = 0.05\,\text{mm})$. The quality team measures samples of $n = 25$ bolts at a time and computes the sample mean diameter $\bar{x}$.
   a. Find $P(\text{individual bolt} > 10.10\,\text{mm})$ using the normal tool at $\mu = 10, \sigma = 0.05$.
   b. Find $P(\bar{x} > 10.10\,\text{mm})$ using the normal tool at the *sampling-distribution* parameters ($\mu = 10$, $\sigma_{\bar{x}} = 0.05/\sqrt{25} = 0.01$).
   c. Which is smaller, and by how much? Why does the same threshold (10.10 mm) become much more "extreme" for $\bar{x}$ than for an individual bolt?
   d. A quality engineer says: "We never see sample means above 10.10 mm — the process is in control." How would you push back if the **individual-bolt** rate of "above 10.10" is 2%?

5. **The normal distribution as a model — not reality.** Heights of adult males are *approximately* normal. Hourly wages in a population are *not*.
   a. Why is **height** well-modeled by a normal — and what feature of the data-generating process makes it so?
   b. Why are **hourly wages** typically *not* well-modeled by a normal? Sketch the shape you'd expect.
   c. Suppose you use a normal model anyway to predict wage percentiles. What kind of errors are you likely to make?
   d. When does the normal still help, even when the underlying variable isn't normal? (Hint: think back to the sampling distributions chapter.)

Chapter 29

Type II Error and Power

Extended Content. This chapter extends beyond a standard one-semester introductory course. It is included for completeness and for instructors who wish to cover additional topics. Content is in draft form.

In the chapter on decision errors we met the two ways a hypothesis test can go wrong and named their probabilities α and β. The Type I error rate α is ours to choose; the Type II error rate β is not — it depends on the unknown truth. This chapter takes a closer look at β: we use simulation to *see* how often a real effect goes undetected, examine the factors that drive the Type II error rate, and connect these ideas to the **power** of a study. This material extends beyond the core STAT 145 sequence and is intended for readers who want to go deeper.

29.1 Recap: the two error types

Recall from the chapter on decision errors:

- A **Type I error** rejects a true H_0; its probability is the discernibility level α, which we choose.
- A **Type II error** fails to reject H_0 when H_A is true; its probability is β.

We can set α directly, but β depends on the effect size, the sample size, and the variability in the data. The rest of this chapter studies what determines β — and how to drive it down.

29.2 Visualizing Type II error

The most effective way to build intuition about Type II error is to think about what happens when there *is* a real effect but our test fails to detect it. Simulation makes this concrete.

29.2.1 A simulation thought experiment

Imagine the following scenario. A tutoring program truly raises students' exam scores by an average of 5 points (on a 100-point scale). The standard deviation of exam scores in the population is 15 points. A researcher recruits $n = 20$ students, has them go through the tutoring program, and compares their scores to the known population mean using a one-sided one-sample t-test ($H_A : \mu > 70$) at $\alpha = 0.05$.

Here is what might happen:

- **The truth**: $\mu = 75$ (population mean with tutoring), while $\mu_0 = 70$ (the null hypothesis value without tutoring). The real effect is $\mu - \mu_0 = 5$ points.
- **One simulated study**: The researcher collects 20 scores. Due to natural variability, the sample mean might be $\bar{x} = 73.2$ with $s = 14.8$. The test statistic is:

$$t = \frac{73.2 - 70}{14.8/\sqrt{20}} = \frac{3.2}{3.31} = 0.97$$

The p-value for this test is about 0.17 — not discernible at $\alpha = 0.05$. The researcher fails to reject H_0, even though the tutoring program really does work. This is a Type II error.

- **Another simulated study**: A different random sample of 20 students happens to include several who benefited greatly from tutoring. The sample mean is $\bar{x} = 76.8$ with $s = 13.5$. Now:

$$t = \frac{76.8 - 70}{13.5/\sqrt{20}} = \frac{6.8}{3.02} = 2.25$$

The p-value is about 0.018 — discernible at $\alpha = 0.05$. This time the researcher correctly rejects H_0. This is a correct decision.

29.2.2 What happens across many studies?

If we repeated this experiment 1,000 times — each time drawing a fresh sample of 20 students from a population where the true effect is 5 points — we would find that roughly 400 of those experiments reject H_0 and roughly 600 do not. That means:

- About 40% of the time, we correctly detect the real effect.
- About 60% of the time, we commit a Type II error.

The Type II error rate is $\beta \approx 0.60$ — alarmingly high! About three in five studies with this design would miss the real effect. This is not because the effect doesn't exist; it's because the sample size is too small relative to the effect size and variability.

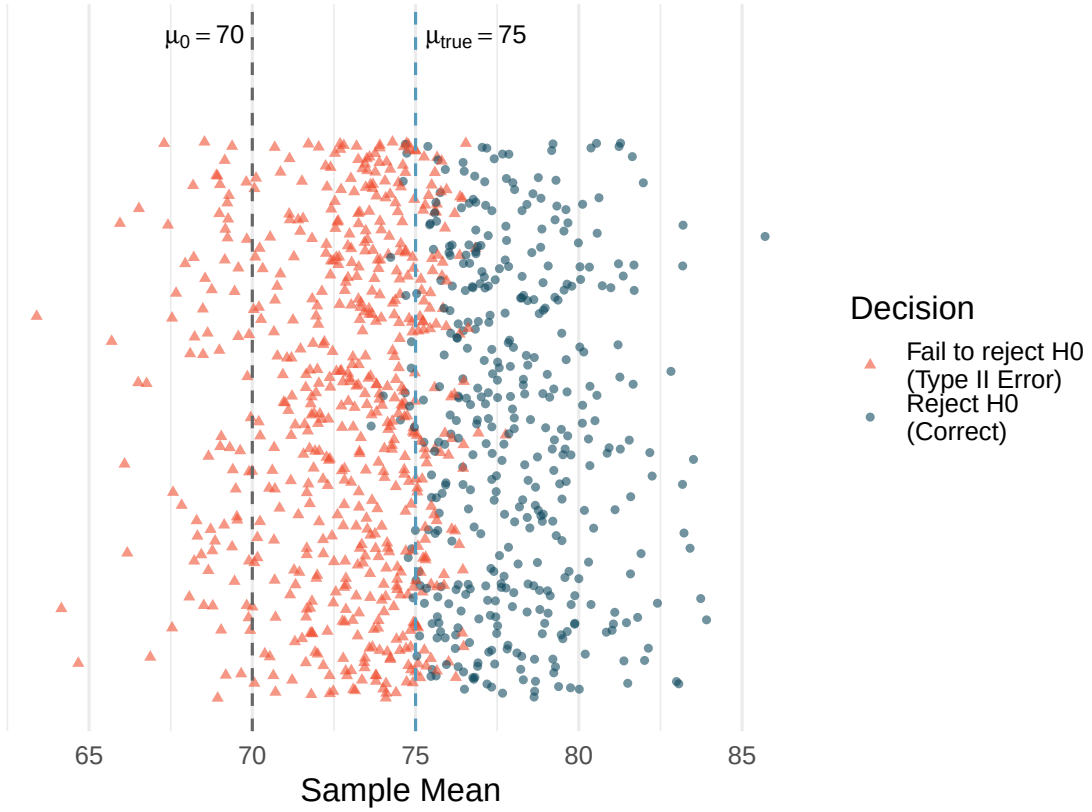


Figure 29.1: Visualization of 1,000 simulated hypothesis tests when the true effect is 5 points. Each point represents one simulated study. Green circles are studies that correctly rejected H□ (correct decisions); red triangles are studies that failed to reject H□ (Type II errors). ↗ Try this live — simulate many studies as hits / Type II misses / Type I false alarms.

In the tutoring program simulation above, what would happen to the Type II error rate if the researcher increased the sample size from 20 to 100 students? Would β increase, decrease, or stay the same?

Show answer

The Type II error rate would decrease substantially. With 100 students instead of 20, the standard error drops from $15/\sqrt{20} = 3.35$ to $15/\sqrt{100} = 1.50$. The test becomes much more precise, making it easier to distinguish the true mean of 75 from the null value of 70. In fact, with $n = 100$, the power would be about 95%, meaning $\beta \approx 0.05$ — the test would miss the real effect only about 5% of the time.

See it in action. Open the Power & Error Visualizer to *see* Type II error: it overlays the null and alternative curves and shades α, β, and the power $1 - \beta$. Increase sample size or effect size and watch β shrink; lower α and watch it grow. For the sim companion, pair with the Power Lab (repeated z-tests, empirical reject-rate).

29.3 Factors that affect the Type II error rate

Four key factors determine how likely a study is to commit a Type II error:

29.3.1 1. Effect size

The **effect size** is the magnitude of the difference between the true parameter value and the null hypothesis value. Larger effects are easier to detect.

- If a tutoring program raises scores by 20 points, almost any reasonable study will detect it.
- If a tutoring program raises scores by 2 points, you need a very large or very precise study to detect it.

Intuitively, a large effect "sticks out" from the noise, while a small effect can easily be hidden by random variability.

29.3.2 2. Sample size

Larger samples provide more precise estimates and smaller standard errors. The standard error of the sample mean is $SE = \sigma/\sqrt{n}$, which decreases as n increases. A smaller standard error means the sampling distribution is narrower, making it easier to distinguish the true effect from zero.

More data, fewer missed effects. Increasing the sample size is the most common way to reduce the Type II error rate. This is why large clinical trials with thousands of participants are more likely to detect treatment effects than small pilot studies.

29.3.3 3. Population variability (σ)

When individual observations are highly variable, it is harder to estimate the population mean precisely. The standard error depends directly on σ: high variability means a wide sampling distribution, which makes it harder to distinguish H_A from H_0.

For example, if exam scores range from 20 to 100 with $\sigma = 25$, it is much harder to detect a 5-point effect than if scores range from 60 to 80 with $\sigma = 5$.

A researcher is studying whether a dietary supplement improves reaction time. She can either recruit participants from the general population (where reaction times are highly variable) or from a group of trained athletes (where reaction times are more consistent). Which group would give her study a lower Type II error rate, assuming the supplement has the same true effect in both populations?

Show answer

The group of trained athletes would give a lower Type II error rate because their reaction times are less variable (smaller σ). A smaller σ leads to a smaller standard error, making it easier to detect the true effect of the supplement.

29.3.4 4. Discernibility level (α)

As we discussed in Section 8.5, a smaller α makes the test more conservative — it demands stronger evidence before rejecting H_0. This directly increases β because the rejection region shrinks.

The table below summarizes these four factors:

Table 29.1: How each factor affects the Type II error rate.

Factor	Change	Effect on β
Effect size	Larger effect	β decreases
Sample size (n)	Larger n	β decreases
Population variability (σ)	Larger σ	β increases
Discernibility level (α)	Smaller α	β increases

A researcher plans a two-sample t-test to compare a new therapy to a standard therapy. She knows from pilot data that the standard deviation of patient outcomes is about $\sigma = 12$. She is hoping to detect a difference of at least 4 points between the two therapies. Her initial plan calls for 30 patients per group with $\alpha = 0.05$.

The standard error for the difference in means is approximately:

$$SE = \sqrt{\frac{12^2}{30} + \frac{12^2}{30}} = \sqrt{\frac{144}{30} + \frac{144}{30}} = \sqrt{9.6} = 3.10$$

The rejection boundary for a two-sided test is approximately $1.96 \times SE = 1.96 \times 3.10 = 6.07$. Since the true effect is only 4 points — smaller than the rejection boundary of 6.07 — the study will more often than not fail to detect the effect. The Type II error rate will be high.

What could she change to reduce the Type II error rate?

She has several options: (1) increase the sample size to reduce the standard error, (2) try to reduce variability through better measurement or a more homogeneous study population, or (3) increase α to 0.10 (though this increases the Type I error rate). The most practical option is usually to increase the sample size.

29.4 Practical vs. statistical discernibility revisited

The concepts of Type I and Type II error are closely connected to the distinction between **statistical discernibility** and **practical importance** that we first encountered in earlier chapters.

Statistical discernibility means the data provide enough evidence to reject H_0 at our chosen α level. It says: "The effect is probably not zero."

Practical importance means the effect is large enough to matter in the real world. It says: "The effect is big enough to be useful."

These are different questions, and they can give different answers:

Scenario	Statistically discernible?	Practically important?	What happened?
Large effect, large n	Yes	Yes	Ideal outcome
Large effect, small n	Possibly not	Yes	Type II error risk — effect exists but may be missed
Small effect, large n	Yes	No	"Discernible but trivial" — very large studies can detect tiny, unimportant effects
Small effect, small n	No	No	No news

The second row (large effect, small n) is the Type II error danger zone: a real, important effect exists, but the study is too small to detect it. The third row (small effect, large n) represents a different problem: a large study that detects a trivially small effect that has no practical importance.

A study with 50,000 participants finds that a new teaching method raises test scores by 0.3 points on a 100-point scale, with $p = 0.002$. Is this result practically important?

The result is statistically discernible ($p = 0.002 < 0.05$), meaning we have strong evidence that the effect is not zero. However, a 0.3-point improvement on a 100-point scale is almost certainly too small to matter in practice. No school would adopt a new teaching method for a gain of one-third of a point. This is a case of statistical discernibility without practical importance — the study was so large that it detected a real but trivially small effect.

Statistical discernibility is not the same as practical importance. A statistically discernible result tells you the effect is probably real, but it does not tell you the effect is important. Always consider the *size* of the effect, not just whether it is statistically discernible.

A small pilot study with 12 participants finds that a new exercise program reduces resting heart rate by 8 beats per minute, but the result is not statistically discernible ($p = 0.09$). Does this mean the exercise program doesn't work?

Show answer

Not necessarily. An 8 bpm reduction in resting heart rate is a clinically meaningful effect. The study may simply have been too small to detect it — this is a potential Type II error. The non-discernible result means we don't have enough evidence to be confident the effect is real, but it doesn't prove the program is ineffective. A larger study with more participants would be warranted.

29.5 Chapter review

29.5.1 Summary

This chapter looked closely at the Type II error rate β and the power of a test. Simulation shows that a study with a small sample, a small true effect, or high variability can fail to detect a real effect a large fraction of the time. Four factors govern β: the **effect size** (larger effects are easier to detect), the **sample size** (more data lowers β), the **population variability** (more variability raises β), and the **discernibility level** (a smaller α raises β). Finally, statistical discernibility and practical importance are not the same: a very large study can detect a trivially small effect, while a small study can miss an important one.

29.6 Exercises

Answers to odd-numbered exercises are provided in the Exercise Solutions appendix at the back of the book.

1. **Which is higher?** In each part below, there is a value of interest and two scenarios: (i) and (ii). For each part, report if the value of interest is larger under scenario (i), scenario (ii), or whether the value is equal under the scenarios.

a. The standard error of $\hat{p}$ when (i) $n = 125$ or (ii) $n = 500$.

b. The margin of error of a confidence interval when the confidence level is (i) 90% or (ii) 80%.

c. The p-value for a Z-statistic of 2.5 calculated based on a (i) sample with $n = 500$ or based on a (ii) sample with $n = 1000$.

d. The probability of making a Type II error when the alternative hypothesis is true and the discernibility level is (i) 0.05 or (ii) 0.10.

2. **Same observation, different sample size.** Suppose you conduct a hypothesis test based on a sample where the sample size is $n = 50$, and arrive at a p-value of 0.08. You then refer back to your notes and discover that you made a careless mistake, the sample size should have been $n = 500$. Will your p-value increase, decrease, or stay the same? Explain.

3. **Practical importance vs. statistical discernibility.** Determine whether the following statement is true or false, and explain your reasoning: "With large sample sizes, even small differences between the null value and the observed point estimate can be statistically discernible."

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **See β on the dual-curve picture.** Open the Power & Error Visualizer.
   a. Where on the picture is β shaded? Why does it live under the **alternative** curve, not the null?
   b. Read the value of β. What does "$\beta = 0.14$" mean in plain English for an experiment at these settings?
   c. Now decrease δ to 0.5 (smaller effect). What happens to β? Why?
   d. Reset and decrease σ to 1 (less noise). What happens to β? Why?
2. **Simulating Type II errors.** Open the Power Lab (simulation).
   a. Run 500 simulated z-tests at these settings. The empirical reject rate is the simulated power. What's the simulated $\beta = 1-$ power?
   b. In the hit/miss/false-alarm strip, what colour represents Type II errors? Roughly what fraction of dots are that colour?
   c. Increase n to 200 and re-run. Does β shrink as expected?
   d. A researcher says: "I'll just call β small because my α is small." Why is this wrong?
3. **When is "no evidence" no evidence?** A trial of $n = 30$ patients tests a new treatment with $H_0 : \mu = 0$ vs. $H_A : \mu > 0$ at $\alpha = 0.05$. The result: p = 0.18, "no discernible effect."
   a. Using the Power & Error Visualizer (preset to $\delta = 0.3, \sigma = 1$), compute the power if the true effect were $\delta = 0.3$. Is that high enough to call "no detected effect" informative?
   b. What if the true effect were $\delta = 1$? Re-check with the tool.
   c. Explain in one sentence the difference between "no evidence of effect" and "evidence of no effect."
   d. What sentence should *actually* appear in the trial's discussion section, given the modest power?

Chapter 30

Statistical Power

Extended Content. This chapter extends beyond a standard one-semester introductory course. It is included for completeness and for instructors who wish to cover additional topics. Content is in draft form.

In the Type II error and power chapter, we saw that a hypothesis test can miss a real effect — a Type II error. The flip side of this coin is **power**: the probability that a test *correctly* detects a real effect. Power is not just a theoretical concept; it is the foundation of good study design. Before collecting data, researchers need to ask: "Is my study large enough to find the effect I'm looking for?" This chapter defines power, explores the factors that influence it, and shows how simulation can estimate power when formulas are unavailable.

30.1 What is power?

In the Type II error and power chapter, we defined the Type II error rate β as the probability of failing to reject H_0 when H_A is true. Power is simply the complement of β.

The **power** of a hypothesis test is the probability of rejecting the null hypothesis when the alternative hypothesis is true:

$$\text{Power} = 1 - \beta = P(\text{reject } H_0 \mid H_A \text{ is true})$$

Power answers the question: "If there is a real effect, how likely is my study to detect it?"

The four possible outcomes of a hypothesis test, with their probabilities, are:

Table 30.1: Probabilities of each outcome in a hypothesis test.

	H_0 is true	H_A is true
Reject H_0	α (Type I error)	$1 - \beta$ (Power)
Fail to reject H_0	$1 - \alpha$ (Correct)	β (Type II error)

Notice the two columns must each sum to 1:

- When H_0 is true: $\alpha + (1 - \alpha) = 1$
- When H_A is true: $(1 - \beta) + \beta = 1$

Power = 1 − β. A test with high power (close to 1) is very likely to detect a real effect. A test with low power (close to 0) will often miss real effects, even when they exist.

30.2 Why power matters

Power matters for both scientific and ethical reasons.

30.2.1 Designing studies that can actually detect effects

Imagine spending a year and $500,000 on a clinical trial, only to discover at the end that the study was too small to detect the treatment effect. The trial produces an inconclusive result — a non-discernible p-value — even though the treatment actually works. The money is gone, the patients' time was wasted, and the scientific question remains unanswered.

This scenario is unfortunately common. A study with low power is essentially a gamble: you might detect the effect (if you're lucky with your sample), or you might not (if natural variability obscures the signal). Low-powered studies are wasteful because they frequently produce null results that leave everyone uncertain.

30.2.2 Ethical considerations

In medical research, every patient enrolled in a clinical trial is exposed to some degree of risk — side effects, time away from proven treatments, or the inconvenience of participation. If a study is underpowered and unlikely to produce a clear answer, then patients are being exposed to risk with little prospect of scientific benefit. Many institutional review boards (IRBs) now require a power analysis before approving research involving human subjects.

30.2.3 The scientific literature

Low-powered studies distort the scientific literature. When only a small fraction of underpowered studies happen to produce discernible results, the published literature overestimates effect sizes (because only the "lucky" studies with unusually large observed effects made it past the discernibility threshold). This phenomenon is sometimes called the **winner's curse** — the studies that "win" (reject H_0) in a low-power setting tend to have inflated effect sizes.

A researcher wants to study whether a new fertilizer increases crop yield. Based on previous research, she expects the fertilizer to increase yield by about 8%. She plans to compare treated and untreated fields. If her study has power of 0.50, what does this mean in practical terms?

Power of 0.50 means that even if the fertilizer truly increases yield by 8%, there is only a 50% chance that her study will detect it — essentially a coin flip. She has a 50% chance of committing a Type II error and concluding there is no evidence the fertilizer works. This is generally considered unacceptable. She should increase her sample size or consider other design improvements before proceeding.

30.3 Factors affecting power

The same four factors that affect β (the Type II error and power chapter) determine power, since power = $1 - \beta$. Here we frame each factor in terms of power rather than Type II error rate.

30.3.1 Sample size: $\uparrow n \rightarrow \uparrow$ power

This is the most important lever researchers have. Larger samples produce smaller standard errors, which make the sampling distribution narrower. A narrower distribution makes it easier to distinguish the alternative from the null.

Power.

The power of the test is the probability of rejecting the null claim when the alternative claim is true.

How easy it is to detect the effect depends on both how big the effect is (e.g., how good the medical treatment is) as well as the sample size.

Consider a clinical trial comparing a new blood pressure medication to a standard one. The new medication is expected to lower blood pressure by 3 mmHg compared to a standard medication. The standard deviation of blood pressure change is $\sigma = 12$ mmHg, and we use a two-sided test with $\alpha = 0.05$.

What is the approximate power with $n = 100$ patients per group? With $n = 500$ per group?

With $n = 100$ per group:

The standard error for the difference in means is:

$$SE = \sqrt{\frac{12^2}{100} + \frac{12^2}{100}} = \sqrt{2.88} \approx 1.70$$

The rejection boundary for a two-sided test at $\alpha = 0.05$ is approximately $\pm 1.96 \times 1.70 = \pm 3.33$ mmHg.

If the true effect is -3 mmHg, then the alternative distribution is centered at -3 with standard deviation 1.70. We need to find the probability that a value from this distribution falls below -3.33:

$$Z = \frac{-3.33 - (-3)}{1.70} = \frac{-0.33}{1.70} = -0.20 \quad \rightarrow \quad P(Z < -0.20) \approx 0.42$$

The power is approximately **42%** — far too low.

With $n = 500$ per group:

$$SE = \sqrt{\frac{12^2}{500} + \frac{12^2}{500}} = \sqrt{0.576} \approx 0.76$$

The rejection boundary is $\pm 1.96 \times 0.76 = \pm 1.49$ mmHg.

$$Z = \frac{-1.49 - (-3)}{0.76} = \frac{1.51}{0.76} = 1.99 \quad \rightarrow \quad P(Z < 1.99) \approx 0.977$$

The power is approximately **97.7%** — excellent. With 500 patients per group, we would almost certainly detect a 3 mmHg effect.

30.3.2 Effect size: ↑ effect →↑ power

A larger effect is easier to detect. If a drug lowers blood pressure by 10 mmHg instead of 3 mmHg, the signal is much stronger relative to the noise, and even a modest sample size will detect it.

Using the same setup as the worked example above ($\sigma = 12$, $\alpha = 0.05$, $n = 100$ per group), estimate the power to detect a true difference of -8 mmHg instead of -3 mmHg.

Show answer

With $SE = 1.70$ and rejection boundary at -3.33, the alternative is centered at -8. We compute $Z = (-3.33 - (-8))/1.70 = 4.67/1.70 = 2.75$, giving $P(Z < 2.75) \approx 0.997$. The power is about 99.7%. An 8 mmHg effect is easy to detect even with only 100 patients per group.

30.3.3 Discernibility level: $\uparrow \alpha \rightarrow \uparrow$ power

A larger α means a wider rejection region. With more "room" to reject H_0, the test is more sensitive — but also more prone to false alarms (Type I errors).

Using the blood pressure setup ($\sigma = 12$, $n = 100$, true effect $= -3$ mmHg), compare the power at $\alpha = 0.05$ versus $\alpha = 0.10$.

Show answer

At $\alpha = 0.05$: rejection boundary $= \pm 1.96 \times 1.70 = \pm 3.33$. $Z = (-3.33 - (-3))/1.70 = -0.20$, power ≈ 0.42.
At $\alpha = 0.10$: rejection boundary $= \pm 1.645 \times 1.70 = \pm 2.80$. $Z = (-2.80 - (-3))/1.70 = 0.12$, power ≈ 0.55.
Power increases from 42% to 55% when we relax α from 0.05 to 0.10, but at the cost of a higher false alarm rate.

30.3.4 Variability: $\uparrow \sigma \rightarrow \downarrow$ power

Greater variability in the data means a larger standard error, which spreads out the sampling distribution and makes it harder to distinguish the alternative from the null.

Table 30.2: Factors that affect power. All relationships assume the other factors are held constant.

Factor	Change	Effect on Power
Sample size (n)	Increase n	Power increases
Effect size	Larger effect	Power increases
Discernibility level (α)	Increase α	Power increases
Population variability (σ)	Decrease σ	Power increases

The researcher's primary tool for increasing power is sample size. Effect size is determined by nature (we don't control how effective a treatment is). Variability can sometimes be reduced through better measurement or study design, but it is largely a feature of the population. The discernibility level α is usually fixed by convention. That leaves sample size as the main adjustable dial.

30.4 Simulation-based power estimation

When exact power formulas are available (as for t-tests and z-tests), we can compute power mathematically. But many real-world situations involve test procedures where exact formulas are complicated or unavailable — for example, bootstrap tests, permutation tests, or tests involving non-normal data. In these cases, **simulation** offers a straightforward alternative.

30.4.1 The simulation recipe

The idea is simple: if we know (or hypothesize) the true state of the world under H_A, we can simulate the data-collection process many times, run the hypothesis test each time, and count how often we reject H_0.

Here is the step-by-step recipe:

1. **Specify the alternative hypothesis**: Choose the effect size you want to detect. For example: "The true mean difference is 5 points."
2. **Specify the data-generating process**: Define the population(s) from which data will be drawn. For example: "Each observation comes from a normal distribution with mean 75 and standard deviation 15."
3. **Choose the sample size** and **discernibility level**.
4. **Simulate one experiment**: Draw a random sample of size n from the specified population, compute the test statistic, and determine whether you reject H_0.
5. **Repeat many times**: Run step 4 a total of B times (e.g., $B = 1{,}000$ or $B = 10{,}000$).

6. **Estimate power**: Count the proportion of simulations that rejected H_0. This proportion is the estimated power.

$$\widehat{\text{Power}} = \frac{\text{Number of simulations that reject } H_0}{B}$$

A biologist plans to test whether a new habitat restoration technique increases the average number of bird species observed at a site. She expects the technique to increase the count from a baseline of 12 species to about 15 species, with a standard deviation of 6 species. She will survey $n = 25$ sites using a one-sample t-test at $\alpha = 0.05$.

Describe how she could use simulation to estimate the power of her study.

---

1. **Alternative**: true mean $= 15$, null value $= 12$.
2. **Data-generating process**: draw 25 observations from a normal distribution with mean 15 and standard deviation 6.
3. **Sample size** $n = 25$, **discernibility level** $\alpha = 0.05$.
4. **One simulated experiment**: generate 25 random values from $N(15, 6)$, compute $\bar{x}$ and s, calculate $t = (\bar{x} - 12)/(s/\sqrt{25})$, find the p-value, and check if $p < 0.05$.
5. **Repeat** 10,000 times.
6. **Estimate power**: in one run of this simulation (`set.seed(30)` in R), 6,706 of the 10,000 simulated studies rejected H_0, so $\widehat{\text{Power}} = 6{,}706/10{,}000 \approx 0.67$.

She would estimate the power at about 67%, short of the conventional 80% target; she would need about 34 sites to reach 80%.

The beauty of simulation-based power analysis is its flexibility. It works for any test procedure — not just t-tests — and can accommodate complex scenarios like non-normal data, unequal group sizes, missing data, or multi-stage designs.

See it in action. Open the Power & Error Visualizer to see the null and alternative sampling distributions side by side with α, β, and power $1 - \beta$ shaded. Drag the effect size, sample size, σ, and α and watch the regions — and the power — change. For the sim view, pair with the Power Lab.

30.5 Power curves

A single power calculation tells us the power for one specific combination of sample size, effect size, α, and σ. But researchers often want to see how power changes across a range of values. **Power curves** show this relationship graphically.

30.5.1 Power vs. sample size

The most common power curve plots power on the y-axis against sample size on the x-axis, for a fixed effect size, α, and σ. This curve answers: "How many observations do I need to achieve a desired level of power?"

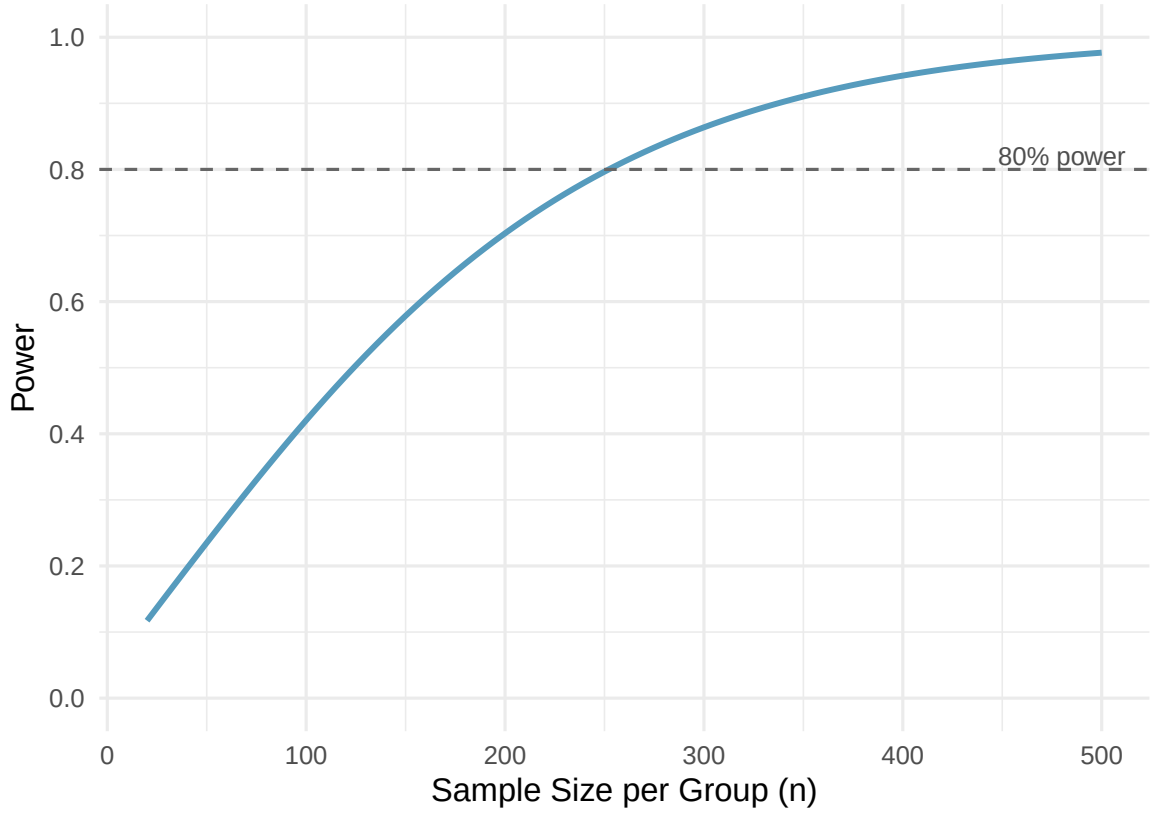


Figure 30.1: A power curve showing power (y-axis) versus sample size per group (x-axis) for the blood pressure example with true effect = 3 mmHg, $\sigma = 12$, $\alpha = 0.05$. The dashed horizontal line marks 80% power; the curve crosses it at approximately n = 250 per group. ↗ Try this live — drag the n slider in the Power & Error Visualizer and trace the curve yourself.

Key features of power curves:

- **Power starts low** for small sample sizes and **increases toward 1** as n grows.
- The curve has a characteristic **S-shape**: power increases slowly at first, then more rapidly in the middle range, then levels off as it approaches 1.
- The curve **never reaches exactly 1** (there is always some chance of a Type II error, though it becomes negligibly small for very large n).
- The steepest part of the curve — where adding more observations gives the biggest payoff in power — is typically in the middle range.

30.5.2 Power vs. effect size

We can also plot power against effect size, holding n, α, and σ fixed. This curve shows how the detectable effect depends on its magnitude.

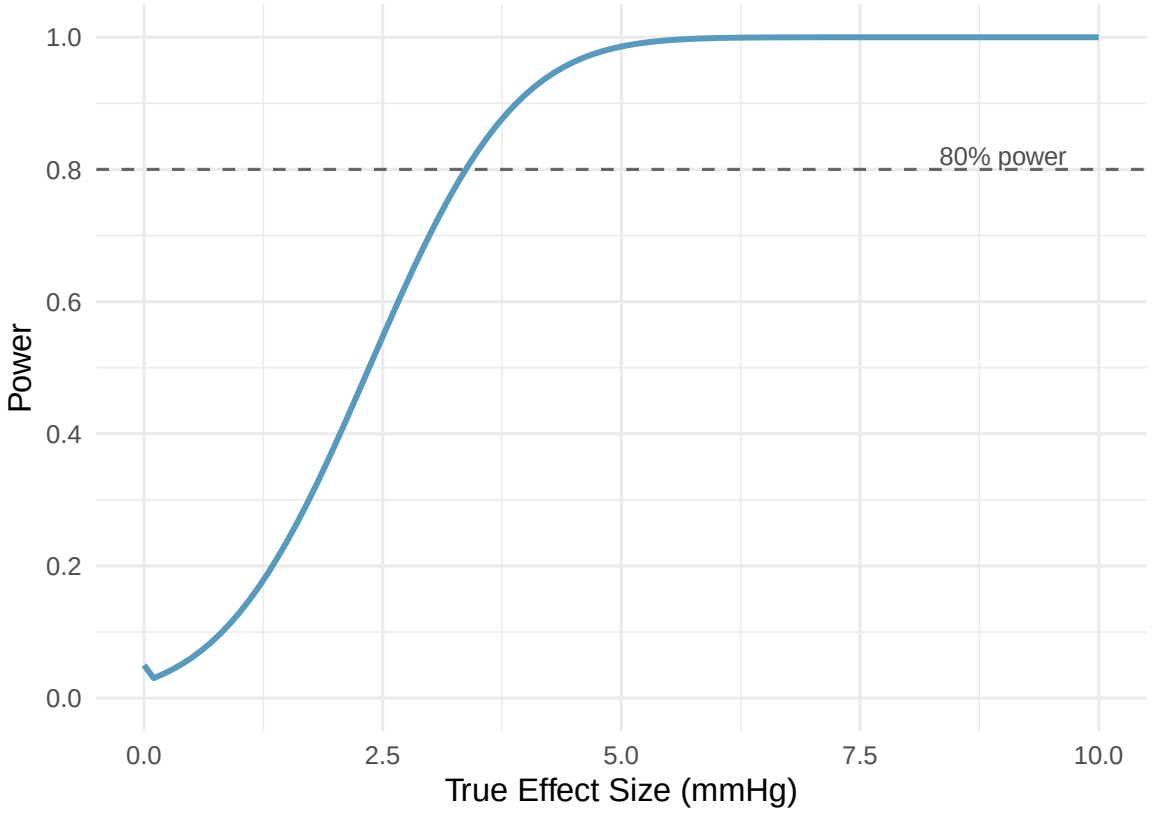


Figure 30.2: A power curve showing power (y-axis) versus true effect size (x-axis) for the blood pressure example with n = 200 per group, $\sigma = 12$, $\alpha = 0.05$. Larger effects are easier to detect. ↗ Try this live — drag the effect-size slider and watch power respond.

At an effect size of 0, the "power" is just α — this is the Type I error rate, the probability of rejecting H_0 when there is no effect. As the effect size grows, power increases.

Look at the power-vs-sample-size curve description above. If a researcher can only afford to recruit 100 patients per group, approximately what power would she have? Is this adequate?

Show answer

From the curve, at $n = 100$ per group the power is approximately 42%. This is far below the conventional minimum of 80%. The researcher should either find a way to increase the sample size, or reconsider whether the study is worth conducting at this scale.

30.6 Typical power targets

How much power is "enough"? There is no universal answer, but strong conventions have emerged in statistical practice.

The 80% power convention. The most widely used target is **power = 0.80** (80%). This means the study has an 80% chance of detecting the specified effect. Equivalently, $\beta = 0.20$ — a 20% chance of a Type II error.

Some fields or funding agencies require higher power:

- **80% power** is the conventional minimum for most research.
- **90% power** is sometimes required for clinical trials, especially when the consequences of a Type II error are severe.

- **95% power** is occasionally used in high-stakes contexts but requires substantially larger sample sizes.

Why is 80% a reasonable target? Why not 95% or 99%?

The choice of 80% reflects a practical tradeoff. Increasing power beyond 80% requires increasingly large sample sizes, and the gains become smaller. Going from 80% to 90% power might require 30-50% more observations. Going from 90% to 95% might require roughly another 25%. At some point, the cost and logistical difficulty of recruiting additional participants outweigh the marginal reduction in Type II error risk.

The 80% target also reflects the observation that in most settings, a Type II error (missing a real effect) is considered less serious than a Type I error (claiming a false effect), so we are comfortable with $\beta = 0.20$ even when $\alpha = 0.05$. If the consequences of missing an effect are severe — for example, in a drug safety trial — a higher power target is appropriate.

Power analysis must be done before data collection. A power analysis conducted *after* data have been collected and analyzed is called **post hoc power analysis** (or "observed power"), and it is widely regarded as misleading. Post hoc power is mathematically determined by the p-value and adds no new information. The time to think about power is during the study design phase.

30.7 Computing power for a two-sample test

While a full derivation of power formulas is beyond the scope of this course, it is useful to see the mechanics of a power calculation for the common case of comparing two group means. This section walks through the logic step by step using the blood pressure medication example.

30.7.1 Setting up the null and alternative distributions

Suppose a pharmaceutical company has developed a new drug to lower blood pressure. They plan a randomized experiment comparing the new drug to a standard medication. The hypotheses are:

- H_0: $\mu_{\text{treatment}} - \mu_{\text{control}} = 0$ (no difference)
- H_A: $\mu_{\text{treatment}} - \mu_{\text{control}} \neq 0$ (there is a difference)

From pilot studies, the standard deviation of blood pressure measurements within each group is about $\sigma = 12$ mmHg. The test will use $\alpha = 0.05$.

Step 1: Determine the null distribution. Under H_0, the difference $\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}$ has mean 0 and standard error:

$$SE = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sigma\sqrt{\frac{2}{n}}$$

With $n = 100$ per group: $SE = 12\sqrt{2/100} = 12 \times 0.1414 = 1.70$.

So under H_0, the test statistic approximately follows a normal distribution centered at 0 with standard deviation 1.70.

Step 2: Find the rejection regions. At $\alpha = 0.05$ (two-sided), we reject H_0 when the observed difference falls more than $1.96 \times SE = 1.96 \times 1.70 = 3.33$ units from zero. That is, we reject when the observed difference is below -3.33 or above $+3.33$.

Step 3: Specify the alternative. Suppose the new drug truly lowers blood pressure by 3 mmHg more than the standard medication, so the true difference is $\mu_{\text{treatment}} - \mu_{\text{control}} = -3$.

Under this alternative, the observed difference follows a normal distribution centered at -3 with standard deviation 1.70 (the standard error doesn't change — it depends on σ and n, not the true mean).

Step 4: Compute the power. Power is the probability that the observed difference falls in the rejection region when the alternative is true. The rejection boundary closest to the true effect of -3 is at -3.33. So we need:

$$P(\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}} < -3.33 \mid \text{true difference} = -3)$$

Standardizing:

$$Z = \frac{-3.33 - (-3)}{1.70} = \frac{-0.33}{1.70} = -0.20$$

$$P(Z < -0.20) \approx 0.42$$

The power is approximately **42%**. This is quite low — the study has less than a coin-flip chance of detecting the 3 mmHg effect.

Using the same setup ($\sigma = 12$, $\alpha = 0.05$, true effect $= -3$ mmHg), compute the power when $n = 250$ per group.

Show answer

With $n = 250$: $SE = 12\sqrt{2/250} = 12 \times 0.0894 = 1.073$. Rejection boundary: $1.96 \times 1.073 = 2.10$. Under H_A centered at -3: $Z = (-2.10 - (-3))/1.073 = 0.90/1.073 = 0.84$. $P(Z < 0.84) \approx 0.80$. The power is approximately 80% — right at the conventional target. So about 250 patients per group would be needed to achieve 80% power for this effect size.

30.8 Chapter review

30.8.1 Summary

Power is the probability of correctly rejecting a false null hypothesis: Power $= 1 - \beta$. A well-designed study should have power of at least 0.80 (80%), meaning it has an 80% chance of detecting the effect of interest.

Power depends on four factors: sample size (more data gives more power), effect size (larger effects are easier to detect), discernibility level (higher α gives more power but more false alarms), and population variability (less variability gives more power). Of these, sample size is the factor most under the researcher's control.

When exact power formulas are unavailable, simulation provides a flexible alternative: generate many datasets under the assumed alternative, run the test on each, and count the proportion that reject H_0. Power curves visualize how power changes with sample size or effect size, helping researchers identify the "sweet spot" where power is adequate without wasting resources.

Power analysis should always be done *before* data collection. Post hoc power analysis — computing power after the study is over — is misleading and widely discouraged.

30.9 Exercises

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Read power off the dual-curve picture.** Open the Power & Error Visualizer.
   a. Identify the two curves on the plot. Which is the null distribution and which is the alternative? How do you tell?
   b. The shaded regions correspond to α, β, and power. Match each label to the correct region.
   c. Read off the current power (the tool reports it as a number). At $\alpha = 0.05$, $n = 30$, $\delta = 1$, $\sigma = 2$, what is it?

   d. In one sentence, explain what "power $= 0.5$" would mean in plain English about your experiment.

2. **The four levers — vary one at a time.** Stay in the Power & Error Visualizer.
   a. Start at $\alpha = 0.05$, $n = 30$, $\delta = 1$, $\sigma = 2$. Note the power. Now double n to 60. How does power change?
   b. Reset to defaults. Double δ (effect size) to 2. How does power change compared to doubling n? Which is the bigger lever?
   c. Reset. Double σ (population noise) to 4. What happens to power? Why?
   d. Reset. Lower α from 0.05 to 0.01. What happens to power? Why is there a tradeoff?

3. **Simulation as a sanity check.** Open the Power Lab (simulation).
   a. Run 500 simulated z-tests under these parameters. The empirical reject rate is the **simulated power**. How close is it to the analytic power from Exercise 1?
   b. Run 5000 simulations and re-read. Tighter agreement?
   c. Look at the dance-of-the-p-values panel. Under a real effect, p-values are *not* uniform — where do they cluster? Why?
   d. A student says: "If the analytic power formula already gives the answer, the simulation is pointless." Give two reasons the simulation is still valuable.

4. **Reading studies from the power perspective.** A journal article reports: "A previous trial ($n = 80$, $\alpha = 0.05$) failed to detect an effect of [the new drug] on blood pressure." A second team plans a replication with $n = 200$.
   a. Why does "the previous trial failed to detect" not mean "the drug doesn't work"?
   b. If the true effect size is small (say $\delta/\sigma = 0.2$), use the Power & Error Visualizer to estimate the power of the first trial. Is the failure to detect surprising?
   c. At what n would the replication achieve at least 80% power for $\delta/\sigma = 0.2$ and $\alpha = 0.05$? Use the tool's n slider.
   d. Suppose the replication achieves power 0.85 and the result is *still* non-significant. What can you conclude — and what can you not?

5. **Power, replication, and "statistical significance."** Suppose 100 independent labs run the same experiment with true power 0.8 at $\alpha = 0.05$.
   a. Roughly how many labs will report a "statistically discernible" result? How many will report a non-result?
   b. Now suppose only the 80 labs with discernible results publish their findings. What kind of bias does this introduce in the literature?
   c. Three independent meta-analyses pool all 100 labs' results (not just the published ones). What advantage does this have over relying on any single significant study?
   d. A media headline reads: "New study shows X causes Y." The study was a single trial with power 0.5. Given everything you've learned, how should you read this headline?

Chapter 31

Sample Size Planning

Extended Content. This chapter extends beyond a standard one-semester introductory course. It is included for completeness and for instructors who wish to cover additional topics. Content is in draft form.

How many observations do you need? This deceptively simple question is one of the most important in all of applied statistics. Too few observations and the study lacks the power to detect real effects (the statistical power chapter). Too many and the study wastes time, money, and — in medical contexts — exposes unnecessary participants to risk. This chapter develops formulas and strategies for determining the right sample size before a study begins.

31.1 Why plan sample size?

Sample size planning — also called **sample size determination** or **power analysis** — is the process of deciding how many observations to collect before beginning a study. There are three main reasons to plan carefully.

31.1.1 Statistical reasons

As we saw in the statistical power chapter, the power of a hypothesis test depends critically on the sample size. A study that is too small has low power and a high probability of committing a Type II error — missing a real effect. A study that is too large may have power close to 100%, but the excess observations provide diminishing returns.

Similarly, for estimation, the width of a confidence interval depends on the sample size through the standard error. More observations produce narrower intervals and more precise estimates. Sample size planning for confidence intervals asks: "How many observations do I need to achieve a desired margin of error?"

31.1.2 Ethical reasons

In medical research, participants are exposed to potential risks: side effects, discomfort, time away from proven treatments. If a study is underpowered and unlikely to produce useful results, participants bear risk without commensurate benefit. Many institutional review boards (IRBs) and research ethics committees require a sample size justification as part of the study approval process.

At the other extreme, an unnecessarily large study exposes more people to risk than needed. If a treatment has a large, harmful side effect, ethical principles demand that we detect it with as few participants as possible.

31.1.3 Practical reasons

Research costs money. Each additional participant means more time, more supplies, more lab work, more data entry. A well-planned sample size ensures that resources are used efficiently — enough to answer the research question, but not so many that the budget is strained or the timeline is unrealistic.

Plan your sample size before collecting data. Determining the sample size after data have been collected defeats the purpose. The goal is to ensure that the study *will have* adequate power or precision, not to verify that it *did*.

31.2 Sample size for estimating a mean

One of the most common sample size questions is: "How many observations do I need to estimate a population mean with a given margin of error?"

31.2.1 The margin of error formula

Recall that a confidence interval for a population mean μ takes the form:

$$\bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$$

where z^* is the critical value for the desired confidence level (e.g., $z^* = 1.96$ for 95% confidence) and σ is the population standard deviation. The **margin of error** (ME) is the "plus-or-minus" part:

$$ME = z^* \cdot \frac{\sigma}{\sqrt{n}}$$

If we want the margin of error to be no larger than some target value, we can solve for n:

$$ME = z^* \cdot \frac{\sigma}{\sqrt{n}} \quad \Longrightarrow \quad \sqrt{n} = \frac{z^* \cdot \sigma}{ME} \quad \Longrightarrow \quad n = \left(\frac{z^* \cdot \sigma}{ME}\right)^2$$

The **sample size for estimating a mean** with margin of error ME at confidence level C is:

$$n = \left(\frac{z^* \cdot \sigma}{ME}\right)^2$$

where z^* is the critical value corresponding to confidence level C (e.g., $z^* = 1.96$ for 95% confidence) and σ is the population standard deviation.

Always round up. When the formula gives a non-integer answer, always round *up* to the next whole number. Rounding down would give a margin of error slightly larger than desired.

A university wants to estimate the average amount of student loan debt among its graduates to within \$1,500 with 95% confidence. Previous studies suggest the standard deviation of student loan debt is about \$12,000. How many graduates should they survey?

We have:

- $ME = 1{,}500$
- $z^* = 1.96$ (for 95% confidence)
- $\sigma = 12{,}000$

$$n = \left(\frac{1.96 \times 12{,}000}{1{,}500}\right)^2 = \left(\frac{23{,}520}{1{,}500}\right)^2 = (15.68)^2 = 245.86$$

Rounding up, the university should survey $n = 246$ **graduates**.

A health researcher wants to estimate the average daily calorie intake of adults in a city to within 50 calories with 99% confidence. From national data, the standard deviation of daily calorie intake is approximately 500 calories. How large a sample does she need?

Show answer

With $ME = 50$, $z^* = 2.576$ (for 99% confidence), and $\sigma = 500$: $n = (2.576 \times 500/50)^2 = (25.76)^2 = 663.6$. Rounding up, she needs $n = 664$ participants.

31.2.2 Where does σ come from?

There is a practical catch in the formula: we need to know the population standard deviation σ before we collect data. But if we already knew σ, we would know more about the population than we actually do. In practice, σ is estimated from:

- **Previous studies**: Published research on similar populations often reports standard deviations that can be used for planning.
- **Pilot studies**: A small preliminary study can provide a rough estimate of σ.
- **Expert judgment**: Subject-matter experts may have reasonable estimates based on experience.
- **Conservative estimates**: When in doubt, use a larger estimate of σ. This produces a larger (more conservative) sample size, ensuring the margin of error target is met even if the true variability is somewhat less than assumed.

Using s from pilot data. If you estimate σ from a pilot study with a small sample, the estimate may be unreliable. When using pilot data, it is wise to inflate the estimate somewhat — for example, using the upper end of a confidence interval for σ — to provide a safety margin.

Referring to the student loan debt example, suppose the \$12,000 standard deviation came from a pilot survey of 30 graduates. The sample standard deviation was $s = 12{,}000$. Should the researcher use exactly $\sigma = 12{,}000$ in the sample size formula?

With only 30 observations, the estimate of σ is itself uncertain. A 95% confidence interval for σ would be roughly \$9,600 to \$16,000. To be safe, the researcher might use $\sigma = 14{,}000$ or even $\sigma = 16{,}000$ in the formula, which would give $n = (1.96 \times 14{,}000/1{,}500)^2 = 335$ or $n = (1.96 \times 16{,}000/1{,}500)^2 = 437.1$, which rounds up to $n = 438$. Using the larger value ensures the margin of error target is likely met.

31.3 Sample size for estimating a proportion

When estimating a population proportion p, the confidence interval takes the form:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The margin of error is:

$$ME = z^* \sqrt{\frac{p^*(1-p^*)}{n}}$$

where p^* is a planning value for p (since we don't know p yet — that's what we're trying to estimate). Solving for n:

$$n = \left(\frac{z^*}{ME}\right)^2 \cdot p^*(1-p^*)$$

The **sample size for estimating a proportion** with margin of error ME at confidence level C is:

$$n = \left(\frac{z^*}{ME}\right)^2 \cdot p^*(1-p^*)$$

where z^* is the critical value for confidence level C and p^* is a planning estimate of the proportion.

31.3.1 Choosing p^*

The value $p^*(1-p^*)$ is largest when $p^* = 0.5$ and smallest when p^* is near 0 or 1. This means:

- If you have a reasonable estimate of p from prior research, use it as p^*.
- If you have no idea what p might be, **use** $p^* = 0.5$. This gives the largest (most conservative) sample size, ensuring the margin of error target is met regardless of the true proportion.

A political polling firm wants to estimate the proportion of voters who support a ballot measure to within 3 percentage points with 95% confidence. They have no prior information about voter support. How many voters should they survey?

With no prior information, we use $p^* = 0.5$ for the most conservative estimate:

- $ME = 0.03$ (3 percentage points)
- $z^* = 1.96$
- $p^* = 0.5$

$$n = \left(\frac{1.96}{0.03}\right)^2 \times 0.5 \times 0.5 = (65.33)^2 \times 0.25 = 4{,}268.4 \times 0.25 = 1{,}067.1$$

Rounding up, they need $n = 1{,}068$ **voters**.

This is why political polls typically survey about 1,000 people — it gives roughly a $\pm 3\%$ margin of error at 95% confidence.

A quality control manager wants to estimate the proportion of defective items in a production run to within 1 percentage point with 95% confidence. Based on historical data, the defect rate is usually around 4%. How large a sample should she inspect?

Show answer

With $ME = 0.01$, $z^* = 1.96$, and $p^* = 0.04$: $n = (1.96/0.01)^2 \times 0.04 \times 0.96 = (196)^2 \times 0.0384 = 38{,}416 \times 0.0384 = 1{,}475.2$. Rounding up, she needs $n = 1{,}476$ items. Note: if she had used $p^* = 0.5$ instead, she would have gotten $n = 9{,}604$ — a much larger and unnecessarily conservative sample. Using prior information about p can save a lot of resources.

Compare the required sample sizes when $p^* = 0.5$ versus $p^* = 0.10$ for a 95% confidence interval with $ME = 0.04$.

With $p^* = 0.5$:

$$n = \left(\frac{1.96}{0.04}\right)^2 \times 0.5 \times 0.5 = (49)^2 \times 0.25 = 2{,}401 \times 0.25 = 600.25 \to n = 601$$

With $p^* = 0.10$:

$$n = \left(\frac{1.96}{0.04}\right)^2 \times 0.10 \times 0.90 = 2{,}401 \times 0.09 = 216.09 \to n = 217$$

Using prior knowledge that p is near 0.10 reduces the required sample from 601 to 217 — a 64% reduction. This illustrates why it is valuable to use a good planning estimate when one is available.

31.4 Sample size for comparing two groups

When comparing two group means, the sample size calculation involves the power framework from the statistical power chapter. The goal is to find n such that the test has at least 80% power (or another target) to detect an effect of a specified size.

31.4.1 Conceptual approach

The idea is the same as in the one-sample case, but now we need to account for two sources of variability (one from each group). For a two-sample t-test comparing means with equal group sizes (n per group), equal standard deviations (σ), and $\alpha = 0.05$ (two-sided):

$$SE = \sigma\sqrt{\frac{2}{n}}$$

To achieve 80% power, we need the true effect size to be about $2.8 \times SE$ away from the null value.[1] Setting the minimum detectable effect δ equal to $2.8 \times SE$ and solving for n:

$$\delta = 2.8 \times \sigma\sqrt{\frac{2}{n}} \quad \Longrightarrow \quad n = 2\left(\frac{2.8 \cdot \sigma}{\delta}\right)^2$$

Returning to the blood pressure example from the statistical power chapter: a new drug is expected to lower blood pressure by at least 3 mmHg compared to a standard medication. The standard deviation is $\sigma = 12$ mmHg. How many patients per group are needed for 80% power with $\alpha = 0.05$?

Using the formula:

$$n = 2\left(\frac{2.8 \times 12}{3}\right)^2 = 2\left(\frac{33.6}{3}\right)^2 = 2 \times (11.2)^2 = 2 \times 125.44 = 250.88$$

Rounding up, $n = 251$ **patients per group** — a total of 502 patients.

This matches the result we found in the statistical power chapter when we computed the power for various sample sizes and found that $n \approx 250$ per group gave approximately 80% power.

A psychologist plans a study comparing reaction times under two conditions. She expects a difference of 25 milliseconds, with a standard deviation of 60 ms in each group. How many participants does she need per group for 80% power at $\alpha = 0.05$?

Show answer

Using the formula: $n = 2(2.8 \times 60/25)^2 = 2(6.72)^2 = 2 \times 45.16 = 90.32$. Rounding up, she needs $n = 91$ per group, or 182 total participants.

31.4.2 Different power targets

The constant 2.8 in the formula above is specific to 80% power and $\alpha = 0.05$ (two-sided). For other targets:

Table 31.1: Multiplier values ($z_{\alpha/2} + z_\beta$) for common power and discernibility level combinations.

Power	$\alpha = 0.05$ (two-sided)	$\alpha = 0.01$ (two-sided)
80%	2.80	3.42
90%	3.24	3.86

[1] The 2.8 comes from $z_{0.025} + z_{0.20} = 1.96 + 0.84 = 2.80$, where $z_{0.025}$ defines the rejection boundary and $z_{0.20}$ defines the 80% power point on the alternative distribution.

Power	$\alpha = 0.05$ (two-sided)	$\alpha = 0.01$ (two-sided)
95%	3.60	4.22

Using Table 31.1, how many patients per group would be needed in the blood pressure example ($\sigma = 12$, $\delta = 3$) for **90% power** at $\alpha = 0.01$?

Show answer

Using the multiplier 3.86: $n = 2(3.86 \times 12/3)^2 = 2(15.44)^2 = 2 \times 238.4 = 476.8$. Rounding up, $n = 477$ per group, or 954 total — nearly double the 502 needed for 80% power at $\alpha = 0.05$. Higher power and stricter α both demand more observations.

31.5 Practical considerations

Formulas give a starting point, but real-world sample size planning involves several additional considerations.

31.5.1 Budget and resources

The formula might say you need 500 participants, but your budget only supports 200. In this case, you have several options:

- **Accept lower power** and acknowledge the study may not detect small effects.
- **Narrow the research question** to focus on larger effects that require fewer observations to detect.
- **Reduce variability** through more precise measurements, a more homogeneous study population, or a paired design (if applicable).
- **Seek additional funding** if the research question is important enough.
- **Don't run the study** if the available sample size gives unacceptably low power. A study that is almost certain to produce null results wastes everyone's time and resources.

Sometimes the best decision is not to run the study. If a power analysis shows that the achievable sample size gives power below 50%, the study is more likely to miss a real effect than to detect one. Researchers should seriously consider whether conducting such a study is worthwhile.

31.5.2 Recruitment and attrition

Not everyone you recruit will complete the study. Participants drop out, miss appointments, provide unusable data, or are lost to follow-up. The sample size formula tells you how many *completed* observations you need. You should plan to recruit more participants to account for expected attrition.

A longitudinal health study requires $n = 300$ completed observations. Based on similar studies, the researchers expect 15% attrition. How many participants should they recruit initially?

If 15% will be lost, then 85% of the recruited participants will complete the study. Let n_{recruit} be the number to recruit:

$$0.85 \times n_{\text{recruit}} = 300 \quad \Longrightarrow \quad n_{\text{recruit}} = \frac{300}{0.85} = 352.9$$

They should recruit **353 participants** to end up with approximately 300 completed observations.

31.5.3 Pilot studies

A **pilot study** is a small preliminary study conducted before the main study. Pilot studies serve several purposes relevant to sample size planning:

- **Estimate** σ (or p) for use in the sample size formula.
- **Test procedures**: Identify problems with the survey instrument, experimental protocol, or data collection process.
- **Estimate attrition rates**: How many participants drop out or provide unusable data?
- **Check feasibility**: Is it realistic to recruit the number of participants the formula requires?

Pilot studies typically involve 10-30 participants — enough to get rough estimates but not so many that significant resources are consumed before the main study begins.

Pilot study data should not be combined with main study data (at least not without careful statistical adjustment). The pilot study is used to *design* the main study. Combining the two can bias the results because the main study design was chosen based on the pilot data.

31.5.4 When formulas don't apply: simulation

The sample size formulas in this chapter assume specific test procedures (z-tests and t-tests) and specific distributional assumptions (normality). When these conditions don't hold — for example, when using a bootstrap or permutation test, or when the data are highly skewed — simulation provides an alternative.

The simulation approach for sample size planning is an extension of the simulation-based power estimation from the statistical power chapter:

1. **Choose a candidate sample size** n.
2. **Simulate** many datasets of size n under the assumed alternative.
3. **Run the analysis** on each simulated dataset.
4. **Compute the proportion** that reject H_0 — this is the estimated power for that n.
5. **Repeat** for different values of n until you find the smallest n that achieves the desired power.

This approach is computationally intensive but extremely flexible. It can handle any test procedure, any data distribution, and any study design.

A researcher wants to use a permutation test (the randomization tests chapter) to compare two groups, but the data are expected to be heavily right-skewed. Formulas for the t-test may not apply. How could she determine the sample size?

She would use simulation:

1. Specify the alternative: for example, Group A has median 50 with a right-skewed distribution, and Group B has median 60 with the same shape.
2. For a candidate n (say, $n = 40$ per group), simulate 10,000 experiments: draw 40 values from each group's assumed distribution, run a permutation test, and record whether H_0 is rejected.
3. Compute the proportion of rejections — this is the estimated power at $n = 40$.
4. Repeat for $n = 50, 60, 80, 100, \ldots$ and find the smallest n where power ≥ 0.80.

This avoids any assumption about normality and accounts for the specific properties of the permutation test.

See it in action. Open the Power & Error Visualizer and slide the sample size n. The null and alternative curves narrow as n grows; the shaded β region shrinks and power $1 - \beta$ climbs. Pair with the Power Lab to confirm empirically: at each n, the simulated reject-rate converges to the analytic power.

31.6 Summary of sample size formulas

The table below collects the key formulas from this chapter for quick reference.

Table 31.2: Summary of sample size formulas. In all cases, round the result up to the next whole number.

Goal	Formula	Key inputs
Estimate a mean to within ME	$n = \left(\frac{z^* \cdot \sigma}{ME}\right)^2$	z^*, σ, ME
Estimate a proportion to within ME	$n = \left(\frac{z^*}{ME}\right)^2 \cdot p^*(1-p^*)$	z^*, p^*, ME
Compare two means (80% power, $\alpha = 0.05$)	$n = 2\left(\frac{2.8 \cdot \sigma}{\delta}\right)^2$ per group	σ, δ (min. detectable effect)

These formulas are starting points, not final answers. Always adjust for attrition, consider practical constraints, and verify assumptions. When in doubt, simulation provides a robust alternative.

31.7 Chapter review

31.7.1 Summary

Sample size planning is a critical step in study design. For estimating a population mean, the required sample size is $n = (z^* \cdot \sigma / ME)^2$, where ME is the desired margin of error. For estimating a proportion, $n = (z^*/ME)^2 \cdot p^*(1-p^*)$, where p^* is a planning value for the proportion (use $p^* = 0.5$ if unknown). For comparing two group means with 80% power, $n = 2(2.8\sigma/\delta)^2$ per group, where δ is the minimum detectable effect.

In all cases, the formula requires prior information about variability (σ or p), which may come from previous studies, pilot data, or conservative estimates. Results should always be rounded up and adjusted for expected attrition.

When standard formulas do not apply — for example, with non-normal data, complex study designs, or non-standard test procedures — simulation-based power analysis offers a flexible alternative. The researcher simulates many datasets under the assumed alternative, runs the planned analysis on each, and identifies the sample size that achieves the desired power.

Practical considerations — budget, recruitment feasibility, attrition, and ethical constraints — always play a role in the final sample size decision. Sometimes the most honest conclusion of a power analysis is that the study cannot be conducted at an adequate scale, and resources should be directed elsewhere.

31.8 Exercises

StatLens Exercises

Let StatLens do the arithmetic; spend your effort on the reasoning a calculator can't do.

1. **Solve for n from a target power.** A clinical trial seeks to detect a small-to-moderate effect ($\delta/\sigma = 0.4$) at $\alpha = 0.05$ (one-sided) with 80% power.
   a. Open the Power & Error Visualizer and slide n up from 30 until the power readout crosses 0.80. What n achieves it?
   b. What n achieves 90% power at the same settings? Why does going from 80% to 90% cost more "extra" n than going from 70% to 80%?
   c. How does the required n change if the true effect is **bigger** ($\delta/\sigma = 0.8$)?
   d. Why is "what n do I need?" a question that **always** requires assuming an effect size? What can a researcher do if she has no good prior estimate?

2. **Sample size for a desired margin of error.** A pollster wants to estimate a population proportion with a 95% CI margin of error no larger than 0.03.
   a. Using the formula $n = (z^*/ME)^2 \cdot p^*(1-p^*)$, what n guarantees ME ≤ 0.03 no matter what p^* turns out to be? (Hint: which p^* maximizes $p^*(1-p^*)$?)
   b. If the pollster knows $p \approx 0.6$, what n achieves ME $= 0.03$? How much smaller than (a)?
   c. Why is using $p^* = 0.5$ a "conservative" (safe) choice in (a)?
   d. A pollster says: "Halving the ME requires doubling n." True or false?
3. **Sample size for two-group comparisons.** A randomized trial compares two treatments and wants to detect a difference of $\delta = 0.5$ standard deviations (a "medium" effect by Cohen's rule) with 80% power at $\alpha = 0.05$ (two-sided).
   a. Using the formula $n_{\text{per group}} = 2 \cdot (2.8 \cdot \sigma/\delta)^2$ (an approximation for $\alpha = 0.05$, 80% power, two-sided), what n per group is required for $\delta/\sigma = 0.5$?
   b. Verify your answer by sliding n on the Power & Error Visualizer until power crosses 0.80.
   c. If attrition is expected to be 20% (1 in 5 participants drops out), how many should you **enroll** to end up with the target n per group?
   d. A study reports "$n = 30$ per group" for a small effect ($\delta/\sigma = 0.3$). What power did this study likely have? Should you trust a non-result?

Foundations — Unit Review

Unit 5 supplies the machinery underneath everything else: the **rules of probability**, the **distributions** that model random outcomes, and the **power/sample-size** ideas that decide whether a study can detect an effect at all. This review connects those pieces and drills two judgments — *which distribution models this situation?* and *what makes a study powerful?*

The big picture

- **Probability** is the language of chance. Probability rules (complements, unions, conditional probability, independence) let you reason about uncertain events, and random variables and expected value summarize a random process by its long-run average and spread.
- **Distributions** are the standard models. The binomial counts **successes in a fixed number of independent yes/no trials**; the normal models **continuous, bell-shaped** measurements — and, via the earlier mathematical-models idea, approximates many statistics in large samples.
- **Power and design.** Type II error, statistical power, and sample size ask the *planning* question: given the choices, how likely is the study to **detect a real effect** — and how many observations does that take?

Which distribution models this?

Situation	Model	You compute
Count of successes in n independent yes/no trials, constant p	**binomial**	$P(X = k)$, mean np, SD $\sqrt{np(1-p)}$
A **continuous, bell-shaped** measurement	**normal**	areas/probabilities via z-scores
A count with **large** n (many successes & failures)	normal **approximation** to the binomial	area under a normal curve

The binomial needs four things: a **fixed** number of trials n, **two** outcomes per trial, **independent** trials, and a **constant** success probability p. If any fails (e.g. sampling without replacement from a small population), the binomial model doesn't apply.

What makes a study powerful?

Power = the probability of detecting a real effect = $1 - \beta$ (where β is the Type II error rate). Power goes **up** when:

1. the **sample size** n is larger,
2. the **true effect** is larger,

3. the **discernibility level** α is larger (a more lenient bar), or
4. the data are **less variable**.

Planning a study is choosing n to reach a target power (often 0.80) for an effect worth detecting — the job of the **sample-size** calculation.

Key ideas to carry forward

Complement / union / intersection · **conditional probability** · **independence** · **random variable** · **expected value** · **binomial** (n, p) · **normal model** · **z-score** · **normal approximation to the binomial** · **Type I error** (α) · **Type II error** (β) · **power** $(1 - \beta)$ · **effect size** · **sample-size planning**.

Common pitfalls

- **Using the binomial when trials aren't independent or** p **isn't constant** (e.g. drawing without replacement from a small group).
- **Confusing** $P(A \mid B)$ **with** $P(B \mid A)$ — conditional probabilities are *directional*.
- **Treating "independent" and "mutually exclusive" as the same** — they're different ideas (mutually exclusive events with nonzero probability are *not* independent).
- **Thinking a non-significant result proves no effect** — it may just reflect **low power** (too small an n).
- **Raising power by raising** α **has a cost** — it increases the Type I error rate.

Exercises

The Unit 5 skills are **choosing the right distribution** and **reasoning about power.** Answers are provided so you can check your work.

1. **Binomial or normal?** For each, name the model you'd use.
   a. The number of heads in 20 fair coin flips.
   b. The heights of adult women, modeled as bell-shaped with a known mean and SD.
   c. The number of defective items in a random sample of 50, each defective with probability 0.03.
   d. A standardized test score, used to find the percentile of a 1300.

   Show answer

   a. binomial $(n = 20, p = 0.5)$. **b.** normal. **c.** binomial $(n = 50, p = 0.03)$. **d.** normal (z-score → percentile).

2. **Do the binomial conditions hold?** You draw 5 cards from a standard deck *without replacement* and count the aces. Is the count binomial? Why or why not?

   Show answer

   No. Without replacement, the probability of an ace **changes** from draw to draw and the trials are **not independent** — two of the binomial requirements fail. (With replacement, it would be binomial.)

3. **What raises power?** A team's study has only 60% power to detect the effect they care about. List three distinct changes that would raise the power, and one cost to watch for.

   Show answer

   Raise power by: **(1)** increasing the **sample size**, **(2)** reducing **variability** (better measurement/design), or **(3)** raising α (a more lenient threshold). Watch the cost: raising α **increases the Type I error rate**, so the usual lever is a larger n. (Detecting a larger true effect also gives more power, but the effect size isn't ours to choose.)

4. **Type I vs. Type II.** A drug trial sets $\alpha = 0.05$. In plain terms, what is the Type I error here, what is the Type II error, and which one does α control?

   Show answer

Type I: concluding the drug works when it truly doesn't (false positive) — its rate is controlled by α (here 0.05). **Type II:** concluding the drug doesn't work when it truly does (false negative) — its rate is β, reduced mainly by a **larger sample**. Power is $1 - \beta$.

5. **Conditional probability.** 1% of a population has a disease. A test is positive in 90% of sick people and in 8% of healthy people. Is $P(\text{disease} \mid \text{positive})$ large or small, and why does intuition often get this wrong?

   Show answer

   It's **small** (roughly 10%): because the disease is rare, the many **false positives** from the 99% healthy population outnumber the true positives. Intuition conflates $P(\text{positive} \mid \text{disease}) = 0.90$ with $P(\text{disease} \mid \text{positive})$ — the **reversed** conditional — which is the base-rate fallacy.

Tech Tutorial: Foundations

Tech Tutorial — Foundations. This unit works with **probability distributions** directly — no data file. You give a distribution its parameters and read off probabilities and percentiles. Pick **one** tool and follow its tab.

To keep things concrete we imagine the *population* behind the class survey: study hours across all intro-stats students are roughly **Normal, mean 13, SD 4**, and a fraction **0.44** of all students live on campus.

What you'll do

1. **Normal** probabilities and percentiles.
2. **Binomial** probabilities.

---

1. Normal distribution

Take study hours to be $X \sim N(\mu = 13,\ \sigma = 4)$. Find:

- (**a**) the probability a student studies **more than 20 hours**, $P(X > 20)$;
- (**b**) the **90th percentile** of study hours.

31.8.1 StatLens

1. Open the Normal Distribution calculator (mean and SD are pre-filled).
2. For (**a**), set the boundary to **20** and select the **upper tail** — read the shaded probability.
3. For (**b**), switch to **inverse / find-x** mode and enter **0.90** to get the percentile.

31.8.2 Jamovi

1. Install the **distrACTION** module (Jamovi library), then **distrACTION → Normal Distribution**.
2. Enter **Mean = 13**, **SD = 4**.
3. Use **Compute probability** with $x_1 = 20$ and the upper tail for (**a**); use **Compute quantile** with $p = 0.90$ for (**b**).

(*Jamovi screenshots to be added.*)

31.8.3 R

```
# (a) P(X > 20) for X ~ N(13, 4)
pnorm(20, mean = 13, sd = 4, lower.tail = FALSE)
```

```
[1] 0.04005916
```

```r
# (b) 90th percentile
qnorm(0.90, mean = 13, sd = 4)
```

```
[1] 18.12621
```

---

2. Binomial distribution

Suppose **44%** of all students live on campus. In a random sample of $n = 12$ students, let X be the number living on campus, so $X \sim \text{Binomial}(n = 12,\ p = 0.44)$. Find:

- **(a)** $P(X = 6)$ — exactly six;
- **(b)** $P(X \geq 6)$ — six or more.

31.8.4 StatLens

1. Open the Binomial Distribution calculator.
2. Set **n = 12** and **p = 0.44**.
3. For **(a)** read the bar at **k = 6**; for **(b)** select the **≥ 6** (upper) region.

31.8.5 Jamovi

1. **distrACTION → Binomial Distribution** (same module as above).
2. Enter **n = 12**, **p = 0.44**.
3. **Compute probability** with $x = 6$ for **(a)**, and the $P(X \geq x)$ option for **(b)**.

(*Jamovi screenshots to be added.*)

31.8.6 R

```r
# (a) P(X = 6) for X ~ Binomial(12, 0.44)
dbinom(6, size = 12, prob = 0.44)
```

```
[1] 0.2067836
```

```r
# (b) P(X >= 6) = 1 - P(X <= 5)
pbinom(5, size = 12, prob = 0.44, lower.tail = FALSE)
```

```
[1] 0.4448014
```

Check yourself

- ☐ find a tail probability and a percentile for a normal distribution;
- ☐ find an exact and a cumulative probability for a binomial distribution;
- ☐ choose the right distribution: normal for a measured amount, binomial for a count of successes in n trials.

> **Going further in R (optional)**
>
> The OpenIntro *IMS* and *OpenIntro Statistics* materials cover the probability foundations; the R functions above (`pnorm`, `qnorm`, `dbinom`, `pbinom`) are the core toolkit. See the normal tutorial for how the normal model underlies later inference.

Chapter A

Technology Setup

Throughout this book you'll use software to do real statistical analysis — making graphs, computing summaries, and carrying out tests and confidence intervals. The *ideas* are the same no matter which tool you use, so the book is built around a shared explanation with a **tabbed choice of tool**: wherever there's a hands-on step, you'll see a **Jamovi** tab and an **R** tab.

> **!** Which tool should I use?
>
> **Your instructor will tell you whether to use Jamovi or R**, and what resources are available to you — for example a shared cloud version, a computer lab, or a Posit Cloud workspace. Pick that tool's tab the first time you see one, and the book will keep showing that tool from then on. You can switch any time.

A.1 Option 1 — Jamovi (point-and-click)

Jamovi is **free, open-source**, and menu-driven — no code required. It's a good fit if you want to focus on the statistics rather than on programming.

- **Install it:** download the desktop version for Windows or Mac from jamovi.org and run the installer.
- **Or use it in a browser:** a cloud version runs at cloud.jamovi.org with nothing to install.
- **Learn it:** the five **Tech Tutorials** (one per unit) and the data **Projects** each carry a Jamovi tab showing exactly which menus to use for that unit's techniques. Pick Jamovi once and your choice follows you through every tabset in the book.

When you open data in Jamovi, your analyses are saved in an `.omv` file that re-runs automatically — so your work is reproducible, and you can always see the underlying steps.

A.2 Option 2 — R (code)

R is a free programming language for statistics, used widely in research and industry. It has a steeper start than Jamovi but gives you a fully reproducible, scriptable workflow.

- **Install it:** install R and then the free RStudio / Posit editor.
- **Or use it in a browser:** Posit Cloud runs RStudio with nothing to install (your instructor may provide a shared workspace).
- **Learn it:** the **R** tabs in those same Tech Tutorials and Projects give runnable code for each technique. To build R skills from scratch alongside them, work through OpenIntro's free, interactive IMS tutorials — self-paced lessons in the same **tidyverse** style this book's code uses.

A.3 Either way: StatLens for the ideas

Separately from Jamovi and R, this book links to **StatLens** — small interactive visualizations you run right in your browser to *see* a concept (a bootstrap distribution forming, a confidence interval's coverage, and so on). StatLens is for building intuition; Jamovi or R is for producing your actual analysis and report.

> **For instructors**
>
> This page is intentionally tool-neutral. Tell students which tab to use and add any institution-specific resources here (shared cloud Jamovi, Posit Cloud workspace links, lab availability, license details).

Chapter B

Datasets

Every dataset used in this textbook is listed below with a one-line description, its source package, and the chapters where it appears. Every dataset is also available as a CSV in this book's `datasets/` folder for use with Jamovi or any other tool.

B.1 From the `openintro` R package

These datasets ship with the openintro R package. Install with `install.packages("openintro")` and access via `openintro::name` or `data(name)` after `library(openintro)`. Each name below links to the package's documentation page.

Dataset	Description	Where used
`antibiotics`	Antibiotic vs. symptomatic treatment for acute sinusitis	Ch 3
`avandia`	Cardiovascular problems in patients on Avandia vs. other diabetes drug	Ch 7
`bac`	Blood-alcohol content vs. beers consumed	Ch 20, 21
`babies_crawl`	Average crawling age vs. birth-month temperature	Ch 19, 21
`bdims`	Body measurements (girths, diameters, sex) of 507 physically active adults	Ch 13, 19, 20
`biontech_adolescents`	Pfizer-BioNTech COVID-19 vaccine trial in 12–15-year-olds	Ch 1
`births14`	1000 births in North Carolina, 2014 (birth weight, weeks, mom's habits)	Ch 13, 14, 20
`china`	Newborn health outcomes vs. rural/urban location in China	Ch 17
`cia_factbook`	259 country-level indicators (life expectancy, GDP, births) from the CIA	Ch 2, 4
`coast_starlight`	Distances and travel times between stops on Amtrak's Coast Starlight	Ch 19, 20
`corr_match`	Scatterplots with known correlations for the match-r-to-plot exercise	Ch 19
`county_2019`	2019 US Census county-level statistics	Ch 20
`county_complete`	Full US county socioeconomic dataset (many years, many variables)	Ch 2, 4

Dataset	Description	Where used
`daycare_fines`	Late-pickup counts at 10 daycare centers before/during/after a fine	Ch 1
`epa2021`	2021 EPA vehicle fuel economy	Ch 14
`exam_grades`	Course grades and exam scores in a college statistics course	Ch 19
`friday`	Traffic and admissions counts on Friday the 6th vs. the 13th	Ch 14
`heart_transplant`	Stanford heart-transplant study — survival by treatment	Ch 3, 7, 12
`hsb2`	High School and Beyond survey subsample (reading, writing, math scores)	Ch 14
`immigration`	Public opinion on immigration policy by age group	Ch 3
`lizard_habitat`	Lizard sun/partial/shade preference by habitat	Ch 16
`lizard_run`	Sprint speed of lizards under two temperature conditions	Ch 14
`loan50`	50 personal loans (income, rate, term, purpose) from Lending Club	Ch 4, 5
`mammals`	Body/brain weight and other traits for 62 mammal species	Ch 5
`mcu_films`	Marvel Cinematic Universe Infinity Saga films (runtimes, gross)	Ch 1
`migraine`	Acupuncture vs. control for migraine pain relief	Ch 1
`nyc_marathon`	NYC Marathon winning times by year and gender	Ch 5
`oscars`	Best Actor and Best Actress Oscar winner ages	Ch 5
`pm25_2022_durham`	Daily PM2.5 air-quality readings, Durham NC 2022	Ch 4
`possum`	Body measurements of 104 mountain brushtail possums	Ch 19, 20, 21
`seattlepets`	Registered pets in Seattle (species, names, breed)	Ch 1
`sinusitis`	Antibiotic vs. symptomatic sinusitis treatment (paired randomized trial)	Ch 1
`smoking`	Smoking status, demographics for 1,691 UK residents	Ch 1
`starbucks`	Nutrition information for Starbucks menu items	Ch 21
`urban_owner`	Homeownership rates by state in urban vs. rural counties	Ch 21
`us_temperature`	US city temperature paired sample (January vs. July)	Ch 14
`yawn`	MythBusters' contagious-yawning experiment	Ch 12

B.2 From other R packages

Dataset	Package	Description	Where used
run17	cherryblossom	2017 Cherry Blossom race times and runner demographics	Ch 1
usairports	airports	US airport locations and characteristics	Ch 1
penguins	palmerpenguins	Palmer Archipelago penguins (species, bill/flipper/mass)	Ch 1
ukbabynames	ukbabynames	UK baby name popularity by year and nation	Ch 1
un_votes	unvotes	UN General Assembly vote records	Ch 1

Install with `install.packages("packagename")`.

B.3 UWL-local datasets

These CSVs are hosted alongside the book and referenced in projects and labs.

Dataset	Description
ai_tool_use.csv	UWL student survey: AI-tool use by age group (small, categorical — good for two-way tables and inference for two proportions)
class_survey.csv	UWL student survey: year, housing, study hours, sleep hours, commute (mixed quantitative + categorical — good for a full EDA project)
flying_etiquette.csv	FiveThirtyEight survey on in-flight etiquette (many categorical variables — rich for chi-square practice)

B.4 Exercise datasets bundled with this book

Four CSVs live in `chapters/data/` rather than in an R package, because the exercises that use them build their figures when the book is rendered. All four are redistributed **unchanged** from *Introduction to Modern Statistics*, which is licensed CC BY-SA 3.0, the same license as this book. The original data sources are listed so you can go back to them.

Dataset	Description	Original source	Where used
meat.csv	Meat and carbohydrate consumption vs. life expectancy in 175 countries, averaged over 2011–2013	You et al. (2022)	Ch 5, 19
netflix_titles.csv	7,788 Netflix titles with type, rating, production country, and release year	TidyTuesday, 2021-04-20, from a Kaggle dataset	Ch 1
stanford-open-policing-combined-data.csv	Traffic-stop summary statistics (stop, search, arrest, and citation rates) by county and driver race	Stanford Open Policing Project (Pierson et al. 2020)	Ch 1

Dataset	Description	Original source	Where used
`survey.csv`	GPA and major for students in an introductory statistics course at Duke University	IMS exercise data	Ch 17

B.5 Using the datasets

In Jamovi or another statistics package. Every dataset above is downloadable as a CSV from the openintro website (e.g., `https://www.openintro.org/data/csv/<name>.csv`), or as tab-separated/txt/rdata from the same page. The UWL-local files are linked directly above.

In R. After installing the source package:

```
library(openintro)
data(penguins, package = "palmerpenguins")  # if the dataset is in another package
head(mcu_films)                              # peek at the first rows
?mcu_films                                   # open the documentation
```

Full openintro dataset catalog. The complete list of openintro datasets (400+, not all used in this book) lives at openintrostat.github.io/openintro/reference/.

> **About this appendix**
>
> The list above is auto-generated by scanning chapter and exercise source files for named dataset references. If you spot a dataset used in the book that isn't listed, please report it via the Report an Issue — Textbook board or annotate the offending chapter in the 145 Textbook Review Hypothesis group.

Chapter C

StatLens Quick Reference

StatLens is a free, browser-based statistics toolkit used throughout this textbook. Every chapter has links to the specific tools it needs.

Browse all tools. The full up-to-date tool index lives at learnlens.org/statlens — distribution calculators, simulation tools, guided activities, and conceptual demos. That is the authoritative catalog. This appendix collects only the two things you may want offline: keyboard shortcuts and the URL parameters used to build reproducible links.

C.1 Keyboard shortcuts

Every StatLens page supports the same keyboard shortcuts. Handy in a lecture demo where you want to advance a simulation without your hand leaving the keyboard.

Key	Action
`1`	Generate +1 sample or resample
`2`	Generate +10
`3`	Generate +100
`4`	Generate +1000
`0`	Reset the simulation
`Space`	Play / pause an animation
`?`	Show the keyboard help overlay
`Esc`	Close a dialog or overlay

C.2 URL parameters

StatLens URLs accept query-string parameters so a link can pre-load a dataset, seed the RNG, or set the confidence level. Instructors use this to build reproducible in-class demos and homework links whose answers are gradeable against a known seed.

Reproducible seeded links

The `seed` parameter uses the same PRNG (sfc32) as StatLens's internal generator, so two viewers loading the same seeded link see the *exact* same sequence of draws. This is what makes seeded homework problems in MyOpenMath auto-gradeable: the assignment engine picks a seed, pre-computes the expected result, then grades the student's answer against it.

Common parameters (not every tool accepts every parameter; each tool's page documents which ones apply):

Parameter	Example	Meaning
`seed`	`?seed=42`	Deterministic PRNG seed
`dataset`	`?dataset=diamonds`	Load a named built-in dataset
`data`	`?data=1,2,3,4,5`	Load inline comma-separated data
`n`	`?n=30`	Set sample size
`p`	`?p=0.5`	Set a probability parameter
`mu`	`?mu=100`	Set a mean
`sigma`	`?sigma=15`	Set a standard deviation
`ci`	`?ci=95`	Set confidence level (percent)
`alt`	`?alt=greater`	Alternative-hypothesis direction (`less`, `greater`, `two-sided`)
`plot`	`?plot=only`	Show only the plot (hide controls) — useful for embedding
`cutlines`	`?cutlines=ci` or `?cutlines=tail`	Draggable cut-lines: CI bounds or tail-area

Combine parameters with &:

`https://learnlens.org/statlens/simulate/bootstrap-mean/?dataset=diamonds&seed=42&ci=95`

For the authoritative parameter list per tool, see the docs at learnlens.org/statlens/docs/url-api.md.

Chapter D

Statistical Tables

Standard reference tables for the four distributions used in this textbook: **standard normal** (z), **Student's** t, **chi-square** (χ^2), and F. Use them for exams or offline work; for everyday computation, use StatLens's distribution calculators, R, or Jamovi.

D.1 Common critical values — cheat sheet

The values below appear most often when constructing confidence intervals or two-sided tests. Read across each row: for a 95% CI or $\alpha = 0.05$ two-sided test, use 1.96 for z^*; use the t^* row for the appropriate df. The last row shows the large-sample limit $t^* \to z^*$.

D.2 Standard normal (z)

Entries are cumulative left-tail probabilities $P(Z \leq z)$ for the standard normal distribution. Read the leading digit and first decimal from the row label, then the second decimal from the column: $z = -1.96$ is row -1.9, column 0.06, and $z = 1.96$ is row $+1.9$, column 0.06. For right-tail probabilities use $1 - P(Z \leq z)$; for two-tailed use symmetry.

Table D.1: Common two-sided critical values (t-star, and z-star in the last row).

df	90%	95%	98%	99%
5	2.015	2.571	3.365	4.032
10	1.812	2.228	2.764	3.169
20	1.725	2.086	2.528	2.845
30	1.697	2.042	2.457	2.750
50	1.676	2.009	2.403	2.678
100	1.660	1.984	2.364	2.626
1000	1.646	1.962	2.330	2.581
∞ (z*)	1.645	1.960	2.326	2.576

Table D.2: Cumulative standard normal probabilities $P(Z \leq z)$ for negative z.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table D.3: Cumulative standard normal probabilities $P(Z \leq z)$ for positive z.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
+0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
+0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
+0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
+0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
+0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
+0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
+0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
+0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
+0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
+1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
+1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
+1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
+1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
+1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
+1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
+1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
+1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
+1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
+1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
+2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
+2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
+2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
+2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
+2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
+2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
+2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
+2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
+2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
+2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
+3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
+3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
+3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
+3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
+3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

D.3 Student's t

Entries are critical values t^* such that $P(T > t^*) = \alpha_{\text{right-tail}}$ for a t-distribution with the row's degrees of freedom. For a two-sided CI at confidence level C, use the column headed $\alpha_{\text{two-tail}} = 1 - C$.

Row: degrees of freedom. **Columns**: one-tail area α to the right of t^*.

> Reading the t-table for a confidence interval
>
> For a two-sided $C\%$ confidence interval, use one-tail area $\alpha = (1 - C/100)/2$. So a 95% CI uses the $\alpha = 0.025$ column; a 99% CI uses the $\alpha = 0.005$ column.

D.4 Chi-square (χ^2)

Entries are critical values χ^{2*} such that $P(X^2 > \chi^{2*}) = \alpha$ for a χ^2-distribution with the row's degrees of freedom.

Row: degrees of freedom. **Columns**: right-tail area α (i.e., $P(X^2 > \chi^{2*}) = \alpha$).

D.5 F-distribution

Entries are critical values F^* such that $P(F > F^*) = \alpha$ for an F-distribution with numerator degrees of freedom df_1 (columns) and denominator degrees of freedom df_2 (rows). Separate tables for $\alpha = 0.05$ and $\alpha = 0.01$.

D.5.1 $\alpha = 0.05$

Rows: denominator df (df_2). **Columns**: numerator df (df_1).

D.5.2 $\alpha = 0.01$

Rows: denominator df (df_2). **Columns**: numerator df (df_1).

D.6 Alternatives

- **StatLens distribution calculators** at learnlens.org/statlens/distribution/ — interactive versions of the same distributions with visual tail-shading.
- **R**: `pnorm`, `qnorm`, `pt`, `qt`, `pchisq`, `qchisq`, `pf`, `qf` (the `p*` functions give left-tail probabilities; the `q*` functions give quantiles/critical values).
- **Jamovi**: use the `descriptives` panel with the `p-value` and `quantile` options, or the distributions module.

Table D.4: Upper-tail critical values of the t-distribution.

df	0.100	0.050	0.025	0.010	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
80	1.292	1.664	1.990	2.374	2.639	3.195
100	1.290	1.660	1.984	2.364	2.626	3.174
200	1.286	1.653	1.972	2.345	2.601	3.131
500	1.283	1.648	1.965	2.334	2.586	3.107
1000	1.282	1.646	1.962	2.330	2.581	3.098

Table D.5: Upper-tail critical values of the chi-square distribution.

df	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Table D.6: F-star with right-tail area 0.05.

df2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	250.10	252.20	253.25
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.48	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.57	8.55
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.69	5.66
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.43	4.40
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.74	3.70
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.30	3.27
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.01	2.97
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.79	2.75
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.62	2.58
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57	2.49	2.45
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.38	2.34
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38	2.30	2.25
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.22	2.18
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.16	2.11
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.19	2.11	2.06
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.15	2.06	2.01
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11	2.02	1.97
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.07	1.98	1.93
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.95	1.90
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.92	1.82	1.77
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.74	1.68
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.74	1.64	1.58
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.53	1.47
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.55	1.43	1.35
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.68	1.58	1.47	1.33	1.24

Table D.7: F-star with right-tail area 0.01.

df2	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6106.32	6157.28	6208.73	6260.65	6313.03	6339.39
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.47	99.48	99.49
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.50	26.32	26.22
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.84	13.65	13.56
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.38	9.20	9.11
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	7.06	6.97
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99	5.82	5.74
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20	5.03	4.95
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65	4.48	4.40
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	4.08	4.00
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	3.94	3.78	3.69
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.70	3.54	3.45
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.51	3.34	3.25
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.35	3.18	3.09
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21	3.05	2.96
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.10	2.93	2.84
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.00	2.83	2.75
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.92	2.75	2.66
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.84	2.67	2.58
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.61	2.52
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.54	2.36	2.27
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.21	2.11
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.20	2.02	1.92
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.84	1.73
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86	1.66	1.53
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.20	2.06	1.90	1.72	1.50	1.35

Chapter E

Exercise Solutions

Solutions to the **odd-numbered** end-of-chapter exercises are provided below, organized by chapter. Even-numbered exercises are reserved for graded assignments; full solutions are available to instructors.

Jump to a chapter

- Ch 1 — Introduction to Data
- Ch 2 — Data Collection and Study Design
- Ch 3 — Exploring Categorical Data
- Ch 4 — Exploring Quantitative Data
- Ch 5 — Relationships Between Variables
- Ch 6 — Sampling Variability
- Ch 7 — Randomization Tests
- Ch 8 — Decision Errors
- Ch 9 — Bootstrap Confidence Intervals
- Ch 10 — Sampling Distributions via Simulation
- Ch 11 — Normal Approximation
- Ch 12 — Inference for Proportions
- Ch 13 — Confidence Intervals for Means
- Ch 14 — Hypothesis Tests for Means
- Ch 15 — Chi-Square Goodness of Fit
- Ch 16 — Chi-Square Test of Independence
- Ch 17 — Analysis of Variance
- Ch 18 — Multiple Comparisons
- Ch 19 — Correlation
- Ch 20 — Linear Regression
- Ch 21 — Prediction and Model Fit
- Ch 22 — ANOVA for Regression
- Ch 23 — Multiple Regression
- Ch 24 — Probability Rules
- Ch 25 — Random Variables
- Ch 26 — Expected Value and Variance
- Ch 27 — Binomial Distribution
- Ch 28 — More About the Normal Distribution
- Ch 29 — Type II Error and Power
- Ch 30 — Statistical Power
- Ch 31 — Sample Size Planning

E.1 Chapter 1 — Introduction to Data

Solutions to the odd-numbered exercises in Chapter 1 — Introduction to Data.

23 observations and 7 variables.

(a) "Is there an association between air pollution exposure and preterm births?" (b) 143,196 births in Southern California between 1989 and 1993. (c) Concentrations of carbon monoxide, nitrogen dioxide, ozone, and particulate matter with an aerodynamic diameter of 10 micrometres or less (PM_{10}) measured at air quality monitoring stations, as well as length of gestation. Continuous numerical variables.

(a) "What is the effect of gamification on learning outcomes compared to traditional teaching methods?" (b) 365 college students taking a statistics course. (c) Gender (categorical), level of studies (categorical, ordinal), academic major (categorical), expertise in English language (categorical, ordinal), use of personal computers and games (categorical, ordinal), treatment group (categorical), score (numerical, discrete).

(a) Treatment: $10/43 = 0.23 \rightarrow 23\%$. (b) Control: $2/46 = 0.04 \rightarrow 4\%$. (c) A higher percentage of patients in the treatment group were pain free 24 hours after receiving acupuncture. (d) It is possible that the observed difference between the two group percentages is due to chance — that is, sampling variability alone could produce a gap this large even if acupuncture had no real effect. (e) Explanatory: whether the patient received acupuncture. Response: whether the patient was pain free after 24 hours.

(a) Experiment; researchers manipulated whether a daycare center imposed a fine and observed how parents' late-pickup behavior changed. (b) 200 cases: the weekly observations of the 10 daycare centers over the 20 weeks. (c) Number of late pickups (numerical, discrete). (d) Week (numerical, discrete), group (categorical, nominal — treatment or control), and study period (categorical, ordinal — before, during, after).

(a) 344 cases (penguins) are included in the data. (b) There are 4 numerical variables: bill length, bill depth, and flipper length (measured in millimeters) and body mass (measured in grams). They are all continuous. (c) There are 3 categorical variables: species (Adelie, Chinstrap, Gentoo), island (Torgersen, Biscoe, Dream), and sex (female, male).

(a) Airport ownership status (public/private), airport usage status (public/private), region (Central, Eastern, Great Lakes, New England, Northwest Mountain, Southern, Southwest, Western Pacific), latitude, and longitude. (b) Airport ownership status: categorical, not ordinal. Airport usage status: categorical, not ordinal. Region: categorical, not ordinal. Latitude: numerical, continuous. Longitude: numerical, continuous.

(a) Year, number of baby girls named Fiona born in that year, and nation. (b) Year (numerical, discrete), number of baby girls named Fiona born in that year (numerical, discrete), and nation (categorical, nominal).

(a) County, state, driver's race, whether the car was searched or not, and whether the driver was arrested or not. (b) All categorical, non-ordinal. (c) Response: whether the car was searched. Explanatory: race of the driver.

(a) Observational study — researchers observed pet-name popularity without assigning any treatment. (b) Dog: Lucy. Cat: Luna. (c) Oliver and Lily. (d) Positive: as the popularity of a name among dogs increases, its popularity among cats tends to increase as well.

E.2 Chapter 2 — Data Collection and Study Design

Solutions to the odd-numbered exercises in Chapter 2 — Data Collection and Study Design.

(a) Population mean, $\mu_{2007} = \$52$; sample mean, $\bar{x}_{2008} = \$58$. (b) Population mean, $\mu_{2001} = 3.37$; sample mean, $\bar{x}_{2012} = 3.59$.

(a) Population: all births in Southern California. Sample: the 143,196 births recorded between 1989 and 1993. (b) If births in this time window at these locations are representative of all Southern California births, the results generalize to that population. Because the study is observational, it cannot establish a causal link between air-pollutant exposure and preterm birth.

(a) The population of interest is all college students studying statistics; the sample is these 365 students. (b) The students were not randomly sampled from the broader population and come from two specific majors, so

generalizing beyond similar students is a stretch. Because the study is an experiment (random assignment to gamification conditions), a causal statement about gamification's effect *within* the studied group is justified.

(a) Observation (a single case). (b) Variable. (c) Sample statistic — a sample mean, $\bar{x}$. (d) Population parameter — a population mean, μ.

(a) Observational. (b) Stratified sampling by section: randomly draw, say, 10 students from each of the four sections for a total sample of 40. This guarantees representation from every section.

(a) Positive, non-linear, moderately strong: countries with higher internet-access rates tend to have higher life expectancy, but the increase levels off around 80 years. (b) Observational. (c) National wealth is a plausible confounder — richer countries can afford both widespread internet access and better health care. (Other reasonable confounders are acceptable.)

(a) Simple random sampling is a reasonable choice — it very rarely goes wrong. (b) Stratifying by field of study is reasonable, since opinions may differ by major. (c) Clustering by age is a poor choice: students of similar ages likely hold similar opinions, so clusters would not be internally diverse (and clusters could be very unequal in size).

(a) The cases are the 200 Mechanical Turk workers who took the survey. (b) Response: attitude toward the fictional microwave oven. (c) Explanatory: dispositional attitude. (d) The 200 participants were recruited randomly online. (e) Observational — participants were not randomly assigned to conditions. (f) No, no causal claim: the study is observational. (g) Generalization to Mechanical Turk workers is reasonable; broader generalization is not.

(a) Simple random sample. Risk: non-response bias — students who reply may hold stronger opinions than those who don't. (b) Convenience sample. Risk: undercoverage bias — friends are unlikely to represent the whole student body; non-response is also possible. (c) Convenience sample — same concerns as (b). (d) Multistage sampling (a random sample of classes, then all students in those classes). If the sampled classes are similar to the population of classes, the main remaining concern is non-response.

(a) Response: exam performance. (b) Explanatory (treatment): light level, with three levels (fluorescent overhead, yellow overhead, no overhead / only desk lamps). (c) Whether the student wears glasses is used as a blocking variable.

(a) Experiment. (b) Two explanatory (treatment) variables: light level (three levels) and noise level (no noise, construction noise, human chatter). (c) Because the researchers want equal representation of glasses-wearers and non-wearers, "wears glasses" is a blocking variable.

Randomize and blind. One workable design: for each participant, prepare two identical cups — one regular Coke, one Diet Coke — labeled A and B, with the A/B $\leftrightarrow$ Coke/Diet assignment randomized independently for each trial. Present the two cups in random order and ask the participant to rate each on how much they liked it. Neither the participant nor the person handing out the cups should know which beverage is which (double-blind). (Other reasonable designs accepted.)

(a) Experiment. (b) Treatment: 25 g of chia seeds twice a day; control: an indistinguishable placebo. (c) Yes — blocking by gender. (d) Single-blind (patients were blinded to their treatment). (e) Because the study is a randomized experiment, causal statements about the studied participants are justified; however, the sample is not random, so the causal claim does not generalize to the broader population.

(a) Non-response bias. The parents who returned the survey are likely those with easier schedules or fewer conflicts; the 85% figure understates the population's difficulty. The school should follow up with non-responders (or take a small random sample of the entire class and pursue high response). (b) The 567 women reached three years later are almost certainly not representative — they are more likely to be homeowners, of higher socioeconomic status, etc. The researcher should compare respondents to the original sample and report the potential bias, not treat the follow-up as a random sample. (c) The study has no comparison group and is observational. People who run may already be healthier, or may exercise in other ways — both are confounders. A stronger design would compare runners to non-runners while controlling for other risk factors, or (better) randomize a running program among willing participants.

(a) Randomized controlled experiment. (b) Explanatory: treatment group (categorical, three levels). Response: psychological well-being. (c) No — participants were student volunteers at one university, so results do not

generalize to the broader population of young adults. (d) Yes — random assignment supports a causal claim about the studied volunteers. (e) Replace "proof" with "evidence": the study provides evidence, not proof.

E.3 Chapter 3 — Exploring Categorical Data

Solutions to the odd-numbered exercises in Chapter 3 — Exploring Categorical Data.

(a) The bar plot shows the order of the categories (by frequency) and lets us read off the relative sizes directly from the bar lengths. (b) There are no features apparent in the pie chart that are not also apparent in the bar plot. (c) The bar plot — lengths are easier for the eye to compare than angles, and the ordering makes the ranking obvious.

(a) Yes, views on the protests and age appear to be associated. The stack breakpoints shift systematically across age groups: older respondents contribute more to the "strongly oppose" segment, while younger respondents lean toward the "strongly support" segment. (b) Answers may vary. Political ideology or party affiliation, education level, and primary news source could all plausibly explain part of the observed pattern.

(a) The stacked (count) plot makes the number of patients in each group easy to see. (b) The standardized (proportion) plot makes the survival rate in each group easy to compare. (c) The standardized plot is the better display for this study — the scientific question is whether transplant improves survival, and comparing proportions between the two groups is exactly what that question asks.

(a) Overall, about 41.0% of JetBlue flights were delayed and 40.7% of United flights were delayed — JetBlue's overall delay rate is slightly higher. (b) By destination: SFO — JetBlue 39.7%, United 40.0%; LAX — JetBlue 40.1%, United 41.0%; BQN — JetBlue 45.7%, United 48.8%. In every city, United's delay rate is higher. (c) This is **Simpson's paradox**. A large share of JetBlue's flights fly to BQN, where delays are high, pulling JetBlue's overall rate up. A large share of United's flights fly to SFO and LAX, where delays are lower, pulling United's overall rate down. The unequal distribution of flights across cities reverses the direction of the comparison when the data are aggregated.

StatLens Exercises

(a) Read the totals directly from the row and column margins of the two-way table; the transplant-and-survived cell is in the interior. A two-way table gives counts *per combination*, not just marginal totals. (b) Row proportions (rows = group) answer "given group, what fraction survived?" The treatment row shows a higher survival rate than the control row — yes, there is an association between receiving a transplant and survival. (c) Column proportions (columns = outcome) answer a *different* question: "given survival, what fraction were in the treatment group?" This is the reverse conditional — useful for a "how many of the survivors received a transplant" question, but not the cause-and-effect question we usually want to answer. Reversing the direction of conditioning changes what the number means. (d) Under independence, both group bars would show the same survived/died split — group membership would be irrelevant to outcome. The actual chart shows the treatment bar with more "survived" and less "died" than the control bar; that mismatch is the visual signature of an association.

E.4 Chapter 4 — Exploring Quantitative Data

Solutions to the odd-numbered exercises in Chapter 4 — Exploring Quantitative Data.

(a) Decrease: the new score is smaller than the mean of the 24 previous scores. (b) Calculate a weighted mean, using weight 24 for the previous mean and weight 1 for the new score: $(24 \times 74 + 1 \times 64)/(24 + 1) = 73.6$. (c) The new score is more than one standard deviation below the previous mean, so the standard deviation increases.

Any 10 employees whose average number of days off is between the minimum and the mean number of days off for the entire workforce at this plant.

(a) Dist B has the higher mean since $20 > 13$, and the higher standard deviation since 20 sits farther from the rest of the values than 13 does. (b) Dist A has the higher mean since $-20 > -40$; Dist B has the higher standard

deviation since -40 is farther from the rest of its data than -20 is from A's. (c) Dist B has the higher mean since every value is larger than the corresponding value in Dist A, but both distributions have the same standard deviation since they are equally variable around their respective means. (d) Both distributions have the same mean (centered at 300), but Dist B has the higher standard deviation since its observations spread farther from the center.

(a) The median is about 26. (b) Since the distribution is right-skewed, the mean is higher than the median. (c) Q_1 is between 15 and 20, Q_3 is between 35 and 40, so the IQR is about 20. (d) Using the $1.5 \times$ IQR rule: upper fence $= Q_3 + 1.5 \times \text{IQR} = 37.5 + 1.5 \times 20 = 67.5$; lower fence $= Q_1 - 1.5 \times \text{IQR} = 17.5 - 1.5 \times 20 = -12.5$. The lowest AQI in the sample is not below 5 and the highest is not above 65, so no day in this sample would be flagged as unusually low or high.

The histogram reveals that the distribution is bimodal, which the box plot hides entirely. The box plot, however, makes it easy to read precise values for observations outside the whiskers.

(a) Right-skewed: there is a natural boundary at 0 and only a few people own many pets. Report the median and IQR. (b) Right-skewed: there is a natural boundary at 0 and only a few people live very far from work. Report the median and IQR. (c) Symmetric. Report the mean and standard deviation. (d) Left-skewed. Report the median and IQR. (e) Left-skewed. Report the median and IQR.

No — we would expect this distribution to be right-skewed, for two reasons: hours of TV watched has a natural lower boundary at 0, and the standard deviation is very large relative to the mean, which forces a long upper tail.

No. The outliers are likely to be the maximum and the minimum themselves, so a statistic based on those values (the range) cannot be robust to outliers.

The 75th percentile score is 82.5, so 5 students will earn an A. By definition, exactly 25% of the class scores above the 75th percentile.

(a) If $\bar{x}/\text{median} = 1$, then $\bar{x} = \text{median}$ — most likely for a symmetric distribution. (b) If $\bar{x}/\text{median} < 1$, then $\bar{x} < \text{median}$ — most likely for a left-skewed distribution, since the mean is pulled down by the low tail more than the median is. (c) If $\bar{x}/\text{median} > 1$, then $\bar{x} > \text{median}$ — most likely for a right-skewed distribution, since the mean is pulled up by the high tail.

(a) About 68% of adults have IQ scores in $(100 - 15,\ 100 + 15) = (85,\ 115)$. (b) About 95% fall in $(100 - 30,\ 100 + 30) = (70,\ 130)$. (c) About 99.7% fall in $(100 - 45,\ 100 + 45) = (55,\ 145)$. (d) About 5% of adults fall outside $(70, 130)$, split evenly between the two tails, so about **2.5%** score at or above 130.

StatLens Exercises

(a) The distribution is **right-skewed and unimodal**: most counties have low poverty rates (roughly 5–15%), with a long tail of higher-poverty counties. (b) With only 5–6 bins, the overall shape (right-skewed) is unchanged, but subtle detail is lost — you can't see whether the peak is smooth or bumpy, and small local clusters get merged. (c) With 30–40 bins, the shape conclusion still doesn't change; extra detail (small local peaks and gaps) becomes visible, but at some point the extra bins add noise rather than signal. (d) The **overall shape** (skew, modality, center, spread) is a real feature of the data and is stable across reasonable binnings. **Fine detail** (exact peak location, minor bumps) depends on the bin width. As long as you don't over- or under-bin dramatically, different histograms of the same data tell the same summary story even if they look different up close.

(a) **Boxplot.** The median appears directly as the line inside the box — read it off. A histogram gives only an approximate median (somewhere near the middle of the mass). (b) **Boxplot.** Extreme values are flagged explicitly as outlier points beyond the whiskers; count how many appear above 5 million. (c) **Histogram** (or dotplot). Skewness is visually obvious — one long tail versus the other. A boxplot shows skew through asymmetric whiskers, but the histogram tells the story faster. (d) With 50+ bins, almost all counties collapse into the leftmost bin because county populations are extremely right-skewed — a few counties (LA, Cook) have millions, but the vast majority have fewer than 100{,}000. The histogram becomes an uninformative single tall spike. A log transformation would help; a boxplot at least shows the outliers explicitly.

(a) Read the mean and standard deviation from the summary panel; NBA heights average roughly 78–79 inches with an SD of about 3–4 inches (exact values depend on the sample). (b) About **68%** of players should fall within

one SD of the mean, by the empirical rule. (c) **Yes** — for a bell-shaped distribution, roughly two-thirds of the data lie inside $(\bar{x} - s,\ \bar{x} + s)$. Not exact, but close. (d) About **95%** within two SDs. Looking at the histogram, nearly all NBA player heights fall inside $(\bar{x} - 2s,\ \bar{x} + 2s)$. (e) **No.** The empirical rule requires an approximately bell-shaped distribution; Ames home prices are strongly right-skewed, so the 68%/95%/99.7% percentages don't apply. Reporting "68% within one SD" would substantially misrepresent the data.

E.5 Chapter 5 — Relationships Between Variables

Solutions to the odd-numbered exercises in Chapter 5 — Relationships Between Variables.

(a) The ridge plots show meat consumption and life expectancy *separately by income group*, not directly against each other. From these plots we can see that the high-income group has both the highest meat consumption and the highest life expectancy, but we cannot see the meat vs. life-expectancy relationship *within* an income group, so we cannot claim they are associated. (b) When a relationship is confounded, we cannot pin down the causal mechanism. A longer life expectancy in high-meat countries could be driven by higher income (which brings better medical care, sanitation, nutrition, and many other life-extending factors) rather than by meat itself. (c) Break the data into subgroups by the confounding variable (income), then examine the relationship of interest (meat vs. life expectancy) separately within each subgroup.

(a) Positive association: mammals with longer gestation periods tend to live longer. (b) The association would still be positive; swapping the axes does not change the direction of a relationship. (c) No, they are not independent — as (a) describes, they trend together.

The plot should show a ramp-up: an early period of nearly exponential growth (few cells reproducing freely), followed by leveling off as the population approaches the one-million-cell carrying capacity. The resulting curve is S-shaped (logistic), rising quickly at first and then flattening out.

Both distributions are unimodal and right-skewed, with a long tail of older winners. The center of the best-actor distribution sits noticeably above the best-actress distribution — best-actor winners are, on average, roughly 7 years older than best-actress winners. Spread is comparable, though the best-actress distribution shows slightly more variability. In short: same shape, similar spread, but best-actor winners tend to be older.

(a) The **manual** transmission group typically shows the higher median city MPG, while the **automatic** group has the larger IQR — the automatic group spans a wider variety of vehicle types (from small commuters to large SUVs), so it is more variable. (b) Both distributions are **unimodal and right-skewed**: most vehicles cluster in the moderate-MPG range, with a long tail of very fuel-efficient outliers (small hybrids and specialty vehicles). (c) The density curves **overlap substantially** in the middle range. The difference between the two group medians is small relative to the within-group variability, so transmission type alone does not explain most of the differences in city MPG. (d) A complete comparison names all four features: **shape** (both right-skewed, unimodal), **center** (manual has slightly higher median), **spread** (automatic has larger IQR), and **unusual features** (a few high-MPG outliers in both groups, likely hybrids or small cars). Overall: a modest difference in center against a backdrop of substantial overlap.

E.6 Chapter 6 — Sampling Variability

Solutions to the odd-numbered exercises in Chapter 6 — Sampling Variability.

(a) Mean — each student reports a numerical value (hours). (b) Mean — each student reports a number (a percentage), and averaging is meaningful. (c) Proportion — each student answers Yes/No, a categorical response. (d) Mean — each student reports a percentage, like in (b). (e) Proportion — each recent graduate reports whether they expect a job within a year, a categorical response.

(a) **The center stays at** μ, the population mean. The sample mean is unbiased, so its long-run average equals μ regardless of n. Bigger samples don't shift the target — they hit it more precisely. (b) **The spread shrinks.** $\text{SE}(\bar{x}) = \sigma/\sqrt{n}$, so quadrupling n from 25 to 100 **cuts the SE roughly in half**; going from 25 to 400 cuts it to about a quarter. (c) **The shape becomes more symmetric and bell-shaped.** At $n = 25$ the sampling distribution can still look lumpy or skewed (depending on the population); by $n = 400$ it typically looks like a smooth mound.

(d) $n = 400$ gives the most precise estimate but requires the most data collection effort. $n = 25$ is cheap but has $4\times$ the SE of $n = 400$. The typical study picks the smallest n that meets the precision target the research question demands.

(a) **Data distribution** — each observation is one patient's age; no averaging across samples. (b) **Sampling distribution** — each observation is the *mean* of a sample of 30, a statistic. (c) **Sampling distribution** — each observation is one person's sample proportion $\hat{p}$; the collection of 400 $\hat{p}$ values is the sampling distribution. (d) **Data distribution** — each observation is one student's SAT score. (e) **The sampling distribution is narrower.** Averaging (or otherwise summarizing) many observations dampens out extremes; $\text{SE}(\bar{x}) = \sigma/\sqrt{n}$, which is smaller than σ once $n > 1$.

(a) $s = 1.4$ **hours.** "A typical student" is about *individual variation* — the sample standard deviation. (b) $\textbf{SE}(\bar{x}) = 0.07$ **hours.** "How much would the *sample mean* vary from sample to sample" is exactly the standard error. (c) s — variability *across students* is the standard deviation of the data. (d) Averaging 400 observations damps out individual variability: $\text{SE}(\bar{x}) = s/\sqrt{n} = 1.4/\sqrt{400} = 0.07$. Dividing by $\sqrt{n} = 20$ shrinks the spread by a factor of 20 relative to s.

Match: **A** $= (n = 100, p = 0.20)$ — centered near 0.20, narrow. **B** $= (n = 25, p = 0.50)$ — centered near 0.50, wide. **C** $= (n = 25, p = 0.20)$ — centered near 0.20, wide. **D** $= (n = 100, p = 0.50)$ — centered near 0.50, narrow. The **center** told you p (which histograms cluster near 0.20 vs. 0.50); the **spread** told you n (wider = smaller n; quadrupling n from 25 to 100 roughly halves the spread).

StatLens Exercises

(a) Population = all 14,000 undergraduates; sample = the 200 surveyed. (b) A **statistic**, $\hat{p} = 0.66$ — computed from the sample. (c) A **parameter**, p — the fraction of *all* 14,000 undergraduates who work 10+ hours. Its value is **unknown**; we would have to survey all 14,000 to know it. The statistic $\hat{p}$ is our best guess at p.

(a) At $n = 5$, the sampling distribution of $\bar{x}$ is *less* skewed than the population but not yet symmetric — still noticeably right-skewed and fairly wide. (Many students predict it will look just like the population; it doesn't.) (b) **Larger** n changed the spread. Drawing *more samples* only fills in the *same* sampling distribution more completely — its spread does not shrink. Increasing n produces a *different, narrower* sampling distribution. These two knobs are easy to confuse. (c) The standard error *is* the standard deviation of the sampling distribution. The observed SD of many simulated sample means estimates that same quantity, so with enough draws the two should nearly agree.

(a) The student confused the **standard error** with the **standard deviation of the data**. An SE of 0.3 describes how much the *sample mean* would vary from sample to sample — not how individual students vary. Individual sleep times vary far more than ± 0.6 hours. (b) The interval $7.1 \pm 2(0.3) = (6.5, 7.7)$ is an approximate **95% confidence interval for the mean** sleep time — a range of plausible values for μ, not a range that captures 95% of individual students.

E.7 Chapter 7 — Randomization Tests

Solutions to the odd-numbered exercises in Chapter 7 — Randomization Tests.

(a) Alternative — claims a directional predictive relationship. (b) Null — "on average run the same speed" asserts no difference. (c) Alternative — claims the probability is *larger* than 0.2. (d) Alternative — claims the mean *has changed*. (e) Null — "risk is equal" asserts no difference. (f) Alternative — claims caffeine *affects* mean birth weight. (g) Null — "the probability is the same" asserts no difference.

(a) In words: H_0: On average, New Yorkers sleep 8 hours per night. H_A: On average, New Yorkers sleep less than 8 hours per night. In symbols: $H_0 : \mu = 8$ vs. $H_A : \mu < 8$, where μ is the mean nightly sleep of New Yorkers. (b) In words: H_0: Employees spend, on average, 15 minutes per business day on non-work activities. H_A: Employees spend more than 15 minutes, on average, on non-work activities during March Madness. In symbols: $H_0 : \mu = 15$ vs. $H_A : \mu > 15$, where μ is the mean non-work minutes per day during March Madness.

(a) (i) **False.** The two groups are very different sizes (67,593 vs. 159,978), so raw counts are not comparable. Compare the *proportions* of cardiovascular events in each group. (ii) **True.** The rate is $2{,}593/67{,}593 \approx 3.8\%$

for Rosiglitazone and $5{,}386/159{,}978 \approx 3.4\%$ for Pioglitazone. (iii) **False.** This is an observational study, so we cannot conclude causation from a difference in rates alone. (iv) **True.** With only the two-way table we cannot separate a genuine treatment effect from ordinary sampling variability. (b) Overall rate of cardiovascular problems: $7{,}979/227{,}571 \approx 0.035$. (c) Under independence we would expect $67{,}593 \times (7{,}979/227{,}571) \approx 2{,}370$ patients in the Rosiglitazone group to experience cardiovascular problems. (d) (i) H_0: Treatment and cardiovascular problems are independent — the observed 3.8% vs. 3.4% gap is due to chance. H_A: The two are not independent — Rosiglitazone is associated with higher risk of cardiovascular problems. (ii) A *higher* proportion in the Rosiglitazone group would give more support for H_A. (iii) The randomization histogram is centered near 0 and its 100 simulated differences never reach the observed difference of about $0.038 - 0.034 = 0.004$; almost none of the shuffles produce a gap this large. That gives strong evidence against independence and supports the alternative — the association between Rosiglitazone and cardiovascular problems is very unlikely to be a chance finding.

StatLens Exercises

(a) The parameter is $\mu_T - \mu_C$, the difference in **population** mean exam scores between the tutoring and control conditions. The observed 4.2-point difference in sample means is our estimate. (b) $H_0 : \mu_T - \mu_C = 0$ vs. $H_A : \mu_T - \mu_C > 0$. The test is **one-sided** because the question asks whether tutoring *improves* scores — a directional claim stated in advance. (c) Under H_0 every student's score would be the same regardless of group label, so the labels are interchangeable. Shuffling the labels generates datasets consistent with that "no-effect" world. **Random assignment** is what makes the shuffle valid: students were placed into groups by chance, not self-selection, so under H_0 the labels really are arbitrary.

(a) The null distribution is centered at **0** — the value H_0 asserts. It must be, because the shuffling procedure breaks any real association between group label and outcome; averaged over many shuffles the differences cancel. (b) The p-value equals (number of shuffled differences at least as extreme as the observed one) / 1000. StatLens reports this directly; for the sex-discrimination data it lands around 0.02–0.03 one-sided. A small p-value means the observed difference is rare in the "no-effect" world, so the data give evidence against H_0. (c) If the observed difference sat near the *middle* of the null distribution, roughly half of the shuffles would be at least that extreme, so the p-value would be near 0.5. We would **fail to reject** H_0 — the observed gap is exactly the kind of fluctuation chance alone routinely produces.

(a) With enough shuffles the null distribution is roughly **bell-shaped**, **centered at 0** by construction, and its **spread** equals the standard error of $\bar{x}_1 - \bar{x}_2$. (b) The standard deviation of the null distribution *is* the simulated standard error of $\bar{x}_1 - \bar{x}_2$. The formula $\text{SE}(\bar{x}_1 - \bar{x}_2) = \sqrt{s_1^2/n_1 + s_2^2/n_2}$ is an analytical estimate of the same quantity, derived under independence and large-sample assumptions. When both are valid, they agree — which is why a randomization test and a t-test on large, well-behaved samples give nearly identical p-values. (c) With small n and an outlier the null distribution is **discrete** (only finitely many label permutations exist), often visibly **non-normal** in the tails, and may show spikes or gaps. The smooth t distribution can over- or under-state the tail probability depending on how the outlier moves around under permutations, so the two p-values can differ substantially. (d) The claim is backwards. The randomization test needs *no* normality assumption — it just enumerates the shuffling distribution that actually applies under random assignment. The t-test relies on the CLT to make $\bar{x}_1 - \bar{x}_2$ approximately normal, and that approximation is *weakest* when n is small. So the randomization test is the *safer* choice in small samples, not the reverse.

E.8 Chapter 8 — Decision Errors

Solutions to the odd-numbered exercises in Chapter 8 — Decision Errors.

(a) H_0: anti-depressants do not affect the symptoms of Fibromyalgia. H_A: anti-depressants do affect the symptoms of Fibromyalgia (either helping or harming). (b) Concluding that anti-depressants either help or worsen Fibromyalgia symptoms when they actually do neither. (c) Concluding that anti-depressants do not affect Fibromyalgia symptoms when they actually do.

(a) **True.** A 99% CI is wider than a 95% CI computed from the same sample (larger t^* or z^*), so it contains every value the 95% interval contains. (b) **False.** Decreasing α *decreases* the Type I error rate — that is what α *is*. Lowering α makes rejection harder and so raises the Type II error rate β at a fixed sample size. (c) **False.** Failing

to reject H_0 does not prove H_0 is true; it only means the evidence against $p = 0.5$ was not strong enough. The true p could be 0.5, or close to 0.5, or the test could simply have lacked power. (d) **True.** With very large n, the standard error becomes tiny, so even a small gap between the null value and the point estimate produces a large test statistic and a small p-value. Statistical discernibility does not imply the effect is practically important.

Seven confidence intervals out of 100 missing π is completely consistent with either a 95% or a 90% confidence level. At 95% we would *expect* about 5 misses in 100 — and 7 is well within the sampling noise you would see even when the students did everything correctly. (At 90% we would expect about 10 misses, so 7 is again unremarkable.) The professor should not dock the seven students for missing π: an occasional CI that misses the parameter is exactly what a confidence *interval* is defined to allow.

In every case, p is the *population* proportion described in the scenario, and H_0 takes an equal sign.

(a) $H_0 : p = 0.20 \quad H_A : p > 0.20$. Here p is the proportion of the 30 pre-test questions a student would answer correctly *in the long run*; random guessing on 5-option questions gives $p = 1/5 = 0.20$, and the professor wants to detect whether students perform better than chance. (b) $H_0 : p = 0.32 \quad H_A : p \neq 0.32$. Here p is the proportion of patients on the new intervention who would experience reduced blood pressure. The trial is looking for a *different* result, not necessarily better, so the alternative is two-sided. (c) $H_0 : p = 0.67 \quad H_A : p > 0.67$. Here p is the proportion of registered voters who will turn out for the next presidential election. The question asks specifically about *higher* turnout, so the alternative is one-sided.

StatLens Exercises

(a) $H_0 : p = 0.50$ versus $H_A : p > 0.50$. (b) **Type I error:** we conclude the drug raises the success rate when it really doesn't. **Type II error:** we conclude the drug doesn't help when it actually does. (c) Patients bear the cost of a Type I error — they receive an ineffective drug (possibly with side effects) in place of a known treatment. Patients and the company bear the cost of a Type II error — a real cure goes undeveloped. When a Type I error has serious safety consequences, regulators demand a small α (often 0.025 or 0.01 one-sided), which raises the evidence bar for rejecting H_0.

(a) About **5%** of the 500 studies reject — by construction, that is what $\alpha = 0.05$ *means*. Every rejection here is a **Type I error** because H_0 is true. (b) About **1%**. The Type I error rate is whatever you set α to (given the test's conditions hold); lowering α from 5% to 1% cuts the false-alarm rate fivefold. (c) About **88–93%** reject — this is the **power** of the test at $p_{\text{true}} = 0.65$, $n = 100$, $\alpha = 0.05$. The non-rejections (roughly 7–12% of studies) are **Type II errors:** H_A is true and we failed to detect it. (d) Power **rises with** n**:** more data makes it easier to distinguish $p = 0.65$ from the null value 0.50. Correspondingly, β falls as n grows.

(a) & (b) At $n = 100$, $\Delta p = 0.15$, and $\alpha = 0.05$ the simulated power is roughly **88–93%** — comfortably above the 80% bar, so this trial *is* adequately powered. A 15-percentage-point effect is large enough that 100 patients detect it most of the time; students often *under*estimate the power available at large effects. (A formal sample-size calculation, coming in Ch 31, gives the precise number.) (c) About a **10% chance** the trial fails to reject H_0 *even though the drug really works* — a Type II error. A non-significant result is never proof the drug does nothing; even well-powered trials carry a residual miss rate. (d) Two levers: **(i) increase** n — larger samples raise power for any fixed effect size and α; this is usually under the researcher's control (budget permitting). **(ii) raise** α — accepting more false positives raises the true-positive rate too; this is usually *not* under the researcher's control once the field has a convention. Other levers include increasing the effect size (enroll patients with more room to improve), decreasing variability (tighter inclusion criteria), or switching to a one-sided test when the direction is truly known in advance.

E.9 Chapter 9 — Bootstrap Confidence Intervals

Solutions to the odd-numbered exercises in Chapter 9 — Bootstrap Confidence Intervals.

(a) The statistic is the observed sample proportion of outdoor videos, $37/128 \approx 0.289$. The parameter is the true proportion of all YouTube videos that take place outdoors — unknown, since we haven't surveyed the entire population. (b) The statistic is $\hat{p}$; the parameter is p. (c) The proportion of "outdoors" videos in each bootstrap resample — a bootstrap sample proportion. (d) Near the observed $\hat{p} \approx 0.289$; a bootstrap distribution is centered at the sample statistic, not at the population parameter. (e) Reading the middle 90% of the histogram gives

roughly $(0.22,\ 0.35)$. (f) We are 90% confident that between about 22% and 35% of all YouTube videos take place outdoors.

With 98% confidence, the true proportion of U.S. adults (in 2022) who get news from social media at least sometimes is between about 0.487 and 0.510. Note that the width is only about 2 percentage points — a consequence of the very large sample size ($n = 12{,}147$).

Match each histogram to the sample proportion it was built from by finding the histogram whose center is nearest that $\hat{p}$ value (a bootstrap distribution is centered at the observed sample statistic). (a) $\hat{p} = 0.13 \rightarrow$ A; (b) $\hat{p} = 0.22 \rightarrow$ B or D (both are near 0.2–0.25); (c) $\hat{p} = 0.30 \rightarrow$ C; (d) $\hat{p} = 0.43 \rightarrow$ has no strong match among these four (the histograms center below 0.35); (e) $\hat{p} = 0.75 \rightarrow$ none.

(a) **Supported.** The entire interval (54%, 64%) lies above 50%, so the data are consistent with a majority. (b) **Not supported.** The value 70% lies above the upper endpoint of 64%, so the interval provides evidence *against* the researcher's conjecture. (c) A 90% CI is *narrower* than the 95% CI — both are centered in the same place, and 70% already sits outside the wider 95% interval, so it will also fall outside the narrower 90% interval. The conjecture would still be rejected.

(a) **Both methods agree.** The SE method assumes an approximately normal bootstrap distribution (statistic ± 2 SE). When the bootstrap distribution is symmetric and bell-shaped, that assumption holds and the two methods give nearly identical intervals. (b) **Percentile is more trustworthy.** The SE method centers the CI symmetrically around $\bar{x}$, but a right-skewed distribution has an asymmetric middle 95% — the upper tail is longer. Percentile captures that asymmetry; SE forces a symmetry that isn't in the data. (c) **Percentile is more trustworthy.** The SE method could push the upper endpoint above 1 (impossible for a proportion). Percentile respects the natural boundary because it just reports what's in the middle 95% of the bootstrap draws. (d) **Neither is fully trustworthy** — a lumpy bootstrap distribution suggests n is too small for the bootstrap to work well. If you must report an interval, percentile is safer than SE because it doesn't require the bootstrap distribution to be roughly normal. (e) The SE method assumes the bootstrap distribution is approximately symmetric and bell-shaped, so that statistic ± 2 SE captures the middle 95%. Percentile makes no shape assumption — it just reads off the 2.5th and 97.5th percentiles. When the bootstrap distribution is skewed, bounded, or lumpy, percentile is the safer default.

StatLens Exercises

(a) The parameter is the *population* median wait time at this drive-through, with symbol M (sometimes $\tilde{\mu}$). The sample median $\tilde{x}$ of the 80 recorded times is the estimate — not the parameter. (b) The mean has a tidy formula for its standard error ($s/\sqrt{n}$) because the CLT gives $\bar{x}$ an approximately normal sampling distribution. The median has no clean closed-form SE, and its sampling distribution can be skewed and irregular. The bootstrap needs no formula — it resamples and *shows* the sampling distribution, so the median works as easily as the mean. (c) Use the **percentile method** when the bootstrap distribution is skewed (a symmetric $\hat{\theta} \pm z^*$ SE would put the interval in the wrong shape). Use the **SE method** when the bootstrap distribution is roughly symmetric and bell-shaped, where the two methods coincide and the SE form is conceptually closer to the formulas that appear later. When in doubt, default to percentile.

(a) Right-skewed, centered near $\hat{p} \approx 0.048$, with narrow spread (small SE). The shape is right-skewed because $\hat{p}$ is small: resamples cannot go below 0 (a hard floor) but can drift somewhat higher, so the tail extends to the right. (b) The 95% percentile interval will land near $(0.01,\ 0.11)$ — typically *just barely* covering the national rate of 0.10. Because the interval contains 0.10, the data do *not* provide strong evidence that this consultant's true complication rate differs from the national average. (c) A symmetric interval $\hat{p} \pm 2\,\text{SE}_{\text{boot}}$ is centered at $\hat{p}$ and the same width on both sides, which is the wrong shape for a right-skewed distribution. Its lower endpoint could even go negative (below 0!), which is impossible for a proportion. The percentile interval respects the skew naturally; the symmetric formula imposes a shape the data don't have.

(a) For well-behaved, moderate-n data the two 95% intervals are typically very close — agreeing to within a small fraction of the SE. The t-procedure's assumptions (independent observations, approximate normality of $\bar{x}$) are met, so the formula and the bootstrap converge. (b) They differ noticeably on small skewed data. The t-interval is forced to be **symmetric** about $\bar{x}$ (equal width on each side, controlled by $t^* \cdot s/\sqrt{n}$); the bootstrap percentile interval can be **asymmetric**, following the actual skew of the bootstrap distribution. The midpoint can shift and the width on one side can exceed the other. (c) The t-interval is built from a *symmetric* distribution (the

t), so it can only express symmetric uncertainty around $\bar{x}$. When the sampling distribution of $\bar{x}$ is itself skewed — which happens with small n and a skewed population — the symmetric interval is the wrong shape, even if its center is right. The bootstrap imposes no shape; it shows what the sampling distribution actually looks like. In well-behaved data the two shapes coincide, so the intervals agree. (d) The intervals aren't "disagreeing about a fact" — they are two approximations to the same target, and they converge as the conditions improve. When they diverge, it is a **signal that the t-interval's symmetry assumption has failed**, not that one of them computed wrong. Trust the **bootstrap** in that case: it makes no normality or symmetry assumption, so it stays honest exactly where the formula breaks down. This mirrors the logic from the proportions chapter: simulation is the more general tool; the formula is the shortcut that only applies when its conditions are met.

E.10 Chapter 10 — Sampling Distributions via Simulation

Solutions to the odd-numbered exercises in Chapter 10 — Sampling Distributions via Simulation.

(a) Sampling distribution of the sample proportion $\hat{p}$. (b) Approximately symmetric. We're told the true proportion is somewhere in $[0.05,\ 0.30]$, and with $n = 800$ the success–failure condition ($np \geq 10$ and $n(1-p) \geq 10$) holds comfortably across that range, so the sampling distribution is bell-shaped. (c) Standard error of $\hat{p}$. (d) With only $n = 250$ homes per sample, the sampling distribution will be **more variable** (wider) than with $n = 800$. Standard error scales as $1/\sqrt{n}$, so shrinking n from 800 to 250 inflates the SE by a factor of $\sqrt{800/250} \approx 1.79$.

(a) **(i) and (iii) are true.** The CLT is a statement about the sampling distribution of $\bar{X}$; it does not change the population distribution or the population mean μ. (b) Mean $= \mu = 3.15$. Standard error $= \sigma/\sqrt{n} = 0.60/\sqrt{400} = 0.030$. (c) The CLT is about the *sampling distribution of* $\bar{X}$ — the distribution of many sample means, one per sample — not about the data in any single sample. A single sample from a skewed population is still skewed no matter how large n is; only the *distribution of means across samples* becomes bell-shaped. (d) A common rule of thumb is $n \geq 30$ for populations that are not too skewed. Strongly skewed or heavy-tailed populations may need $n = 100$ or more. There is no single magic number: the rougher the population, the larger n needs to be.

StatLens Exercises

(a) Most students predict (ii). (b) In practice you'll see (iii): at $n = 5$ from a strongly right-skewed population, the sampling distribution of $\bar{x}$ still carries visible skew — the CLT hasn't kicked in yet. (c) By $n = 30$ the skew is usually mostly gone for a moderately skewed population; by $n = 100$ the distribution looks essentially normal. (d) A **bimodal** population typically needs a *larger* n than a right-skewed one, because the two modes have to "average out" before the bell shape emerges. The CLT still eventually wins, but "$n = 30$" is a rule of thumb, not a guarantee — the shape of the population dictates how fast the sampling distribution becomes normal.

(a) Predicted center $= p = 0.20$. Predicted spread $= \sqrt{p(1-p)/n} = \sqrt{(0.20)(0.80)/20} = \sqrt{0.008} \approx 0.089$. (b) With 5000 samples the lab should report a center very close to 0.20 and an SD of $\hat{p}$ very close to 0.089. (c) **Right-skewed.** With $p = 0.20$ and $n = 20$, $\hat{p}$ has a hard floor at 0 and a long tail toward larger values; the success–failure check gives $np = 4 < 10$, so the normal approximation is poor here. (d) At $n = 200$, $np = 40$ and $n(1-p) = 160$ — both well above 10 — so the sampling distribution is now approximately normal and much tighter, with SE $= \sqrt{(0.20)(0.80)/200} \approx 0.028$ (about one-third the earlier spread).

(a) Students at one table are **not independent** of each other: they're likely clustered in friend groups, dietary preferences, or ordering patterns. The CLT's SE formula assumes i.i.d. observations, so it *underestimates* the actual variability of $\bar{x}$; a CLT-based CI will under-cover the true mean (real coverage far below 95%). (b) The Cauchy distribution is the classic CLT counterexample: it has no defined mean or variance, and the sampling distribution of $\bar{x}$ **stays Cauchy at every** n — it never becomes normal, no matter how large n gets. (c) Truly infinite-variance populations are rare in practice; subtle clustering (same dorm, same city, same day) is everywhere. **Independence is the assumption to worry about**, not the existence of finite variance. (d) Not very comforting. When the distribution is heavy-tailed enough that a single viral post can drive $\bar{x}$, the CLT still applies "in the limit" but the limit is approached very slowly — the practical n needed for normality may be in the millions, not the thousands.

E.11 Chapter 11 — Normal Approximation

Solutions to the odd-numbered exercises in Chapter 11 — Normal Approximation.

(a) The general formula is point estimate $\pm\, z^\star \times SE$, with point estimate $= 45\%$, $z^\star = 1.96$ for 95% confidence, and $SE = 1.2\%$: $45\% \pm 1.96 \times 1.2\% \to (42.6\%, 47.4\%)$. We are 95% confident that the proportion of U.S. adults who live with one or more chronic conditions is between 42.6% and 47.4%. (b) (i) **False.** A 95% confidence interval will "miss" the true value about 5% of the time. (ii) **True.** Long-run coverage: about 950 of 1,000 such intervals would capture the true proportion. (iii) **True.** 50% lies outside the interval, so the null value $p = 0.5$ would be rejected at $\alpha = 0.05$. (iv) **False.** Standard error describes the sampling variability of the estimate, not individuals' uncertainty about their own answers.

(c) is the correct interpretation: **a Z score of 0.47 means the sample proportion is 0.47 standard errors greater than the hypothesized value of the population proportion.** The others describe different ideas: (a) confuses the Z score with a probability that H_0 is true; (b) describes a *p-value*, not a Z score; (d) confuses the sample proportion itself with a scaled distance; (e) drops the units — it's "0.47 SEs away," not "0.47 units away"; (f) claims $\hat{p} = 0.47$, which is unrelated.

(a) $P(Z < -1.35) \approx 0.0885$, or 8.85%. (b) $P(Z > 1.48) \approx 0.0694$, or 6.94%. (c) $P(-0.4 < Z < 1.5) = P(Z < 1.5) - P(Z < -0.4) \approx 0.9332 - 0.3446 = 0.5886$, or about 58.9%. (d) $P(|Z| > 2) = P(Z < -2) + P(Z > 2) \approx 2(0.0228) = 0.0455$, or about 4.6%.

(a) Verbal Reasoning: $N(\mu = 151, \sigma = 7)$; Quantitative Reasoning: $N(\mu = 153, \sigma = 7.67)$. (b) $Z_{VR} = (160 - 151)/7 = 1.29$; $Z_{QR} = (157 - 153)/7.67 = 0.52$. (c) Sophia scored 1.29 standard deviations above the mean on Verbal Reasoning and 0.52 standard deviations above the mean on Quantitative Reasoning. (d) She did better on Verbal Reasoning — her Z score is higher there. (e) $Perc_{VR} \approx 0.90$ (90th percentile); $Perc_{QR} \approx 0.70$ (70th percentile). (f) About 10% of test takers did better than Sophia on Verbal Reasoning and about 30% did better on Quantitative Reasoning. (g) Raw scores are on different scales (different means and SDs), so a raw comparison is misleading; percentiles put both sections on a common footing. (h) The Z scores in part (b) would not change (a Z score is just a rescaling), but parts (d)–(f) would no longer be answerable — percentiles require a distributional model.

(a) The 80th percentile of $N(153, 7.67)$: $z = 0.84$, so the score is $153 + 0.84 \times 7.67 \approx 159$. (b) "Worse than 70% of test takers" is the 30th percentile of $N(151, 7)$: $z = -0.52$, so the score is $151 + (-0.52) \times 7 \approx 147$.

(a) $Z = (83 - 77)/5 = 1.2$, and $P(Z > 1.2) \approx 0.1151$ — about an 11.5% chance of an 83°F day or hotter. (b) The 10th percentile: $z = -1.28$, so the cutoff is $77 + (-1.28)(5) \approx 70.6$°F. The coldest 10% of June days in LA have highs of about 70.6°F or lower.

(a) Converting via $C = (F - 32) \cdot 5/9$: the mean becomes $(77 - 32) \cdot 5/9 = 25$°C and the SD scales the same way, $5 \cdot 5/9 \approx 2.78$°C. So $N(25, 2.78)$. (b) $Z = (28 - 25)/2.78 \approx 1.08$; $P(Z > 1.08) \approx 0.1401$. (c) The answer is very close to part (a) of the previous exercise (about 11.5%). Not surprising: only the units changed. The tiny difference is because 28°C corresponds to 82.4°F, not exactly 83°F. (d) Using $Q_1 = 25 + (-0.674)(2.78) \approx 23.13$°C and $Q_3 = 25 + (0.674)(2.78) \approx 26.87$°C, the IQR is ≈ 3.73°C.

(a) **Panel C ($n = 200$).** Larger n makes the sampling distribution smoother, more symmetric, and more sharply peaked — the normal curve hugs it closely. (b) The $n = 10$ histogram is visibly right-skewed and bumps into the lower boundary at 0. The **success–failure condition** fails: $np = 10(0.30) = 3 < 10$. (c) Panel A: $np = 3$, $n(1-p) = 7$ — **fails.** Panel B: $np = 15$, $n(1-p) = 35$ — **passes.** Panel C: $np = 60$, $n(1-p) = 140$ — **passes clearly.** (d) The normal approximation is useful because *sampling distributions* (not raw data, not populations) tend toward normal as n grows. It requires enough sample size for skewness and discreteness to smooth out; when the success–failure condition fails, prefer a simulation approach.

StatLens Exercises

(a) Area to the right of $z = 1$ is about **0.159**; the area to the left of $z = -1$ is also **0.159**. They're equal because the standard normal is **symmetric about 0**. (b) About **0.954** for $-2 \le z \le 2$ — consistent with "95%" within ± 2 SDs (the empirical rule's ± 2 figure is really ± 1.96; both round to 95%). (c) $z \approx -1.96$. This is the **critical value** for a 95% confidence interval; you'll use $z^\star = 1.96$ throughout the CI and two-sided z-test material at $\alpha = 0.05$. (d) Total area under any density curve is **1** by definition (the value is somewhere). Area = 0 corresponds to an event

like "z equals exactly 0.5" — continuous distributions assign probability 0 to any single point; only intervals carry positive probability.

(a) Mean $= 520$, standard error $= 110/\sqrt{25} = 22$, shape approximately normal. (b) $z = (550 - 520)/22 \approx 1.36$, so $P(\bar{x} \geq 550) \approx 0.087$. (c) For an individual student, $z = (550 - 520)/110 \approx 0.27$, so $P(X \geq 550) \approx 0.39$ — **much bigger.** Individual scores vary a lot; sample means vary less because averaging dampens individual variability, making the same 30-point gap "more surprising" for a mean than for one student. (d) **Smaller.** Doubling n shrinks the SE to $110/\sqrt{50} \approx 15.6$. The same 30-point gap becomes $z \approx 1.93$ (instead of 1.36) — farther in the tail, lower probability.

(a) For well-behaved data ($\hat{p}$ near 0.5, n large) the two 95% intervals are very close — typically agreeing to within a fraction of a percentage point. (b) They differ noticeably. The z-interval is **symmetric** about $\hat{p}$; the bootstrap interval is **asymmetric**, following the shape of the bootstrap distribution. (c) At small $\hat{p}$ the bootstrap distribution is **right-skewed** — $\hat{p}$ has a hard floor at 0, but resamples can drift upward. The z-interval forces symmetry that doesn't match. (d) The two intervals are not both "right." The z-interval **under-covers** in the troublesome case because its symmetry assumption fails. The bootstrap is more **honest**: its long-run coverage matches the advertised 95% because it reflects the actual shape of the uncertainty. "Honest" here means the interval's real coverage matches its stated confidence level.

E.12 Chapter 12 — Inference for Proportions

Solutions to the odd-numbered exercises in Chapter 12 — Inference for Proportions.

First, the hypotheses should be about the population proportion (p), not the sample proportion $\hat{p}$. Second, the null value should be the claimed value (0.25), not the observed value (0.29). Correct setup: $H_0 : p = 0.25$ and $H_A : p > 0.25$.

(a) $H_0 : p = 0.20$, $H_A : p > 0.20$. (b) $\hat{p} = 159/650 = 0.245$. (c) Represent each respondent with a card: use 100 cards, 20 black (support "defund") and 80 red (do not). Shuffle and draw 650 cards *with replacement*, calculate $\hat{p}_{\text{sim}}$, and repeat many times (using software in practice). The simulated null distribution shows what $\hat{p}$ values are typical if $p = 0.20$. The p-value is the proportion of simulations with $\hat{p}_{\text{sim}} \geq 0.245$. (d) The p-value is approximately 0.001 (only about 1 in 1,000 simulated proportions reaches 0.245 or higher). Because p-value < 0.05, we reject H_0: the data provide convincing evidence that more than 20% of Seattle adults support proposals to defund police departments.

(a) $H_0 : p = 0.5$, $H_A : p \neq 0.5$. (b) The p-value is roughly 0.4. There is not enough evidence to conclude that cats prefer one shape over the other — but note that with only 7 cats there is very little power to detect a preference even if one exists.

(a) $SE(\hat{p}) \approx 0.189$. (b) Roughly 0.188 (matches the theoretical SE closely). (c) Yes — the simulation-based null distribution and the normal model agree. (d) No — with only $n = 7$ cats, the simulated $\hat{p}$ values are visibly *discrete* (only a handful of distinct outcomes are possible). (e) The bootstrap draws are discrete (only a few distinct proportions occur for such a small sample), while the mathematical normal model is continuous.

(a) The null hypothesis distribution assumes $p = 0.7$; the bootstrap distribution is built from the data, which had $\hat{p} = 0.6$. (b) The null distribution is centered at 0.7; the bootstrap distribution is centered at 0.6. (c) Both distributions have a standard error of roughly 0.1. (d) Both are reasonably symmetric. The null distribution is slightly more skewed (left) than the bootstrap because 0.7 is closer to the boundary at 1, and the boundary compresses the right tail.

(a) The null-hypothesis simulation is used for hypothesis testing; the data-bootstrap distribution is used for confidence intervals. (b) $H_0 : p = 0.7$ vs. $H_A : p \neq 0.7$. The p-value > 0.05, so there is no evidence that the proportion of full-time statistics majors who work at least 5 hours per week differs from 70%. (c) We are 98% confident that the true proportion of full-time statistics majors who work at least 5 hours per week is between 35% and 80%. (d) Using $z^\star = 2.33$, the 98% CI is approximately $(0.367, 0.833)$.

(a) False. The success-failure condition fails (only 8% of a small sample), so the sampling distribution will *not* be approximately normal. (b) True. Because $p = 0.08$ is close to 0 and $\hat{p}$ cannot go below 0, the sampling distribution is right-skewed for small samples. (c) False. $SE_{\hat{p}} = 0.0243$, and $\hat{p} = 0.12$ is $(0.12 - 0.08)/0.0243 \approx$

1.65 SEs away — not unusual. (d) True. $\hat{p} = 0.12$ is 2.32 SEs from p, which is unusual. (e) False. Doubling n decreases the SE by a factor of $1/\sqrt{2}$, not $1/2$.

(a) False. A confidence interval estimates the *population* proportion, not the sample proportion. (b) True. A 95% CI is approximately $82\% \pm 2\%$. (c) True. This is what the confidence *level* means: about 95% of intervals constructed this way in repeated sampling will capture p. (d) True. Quadrupling n shrinks the margin of error by a factor of $1/\sqrt{4} = 1/2$. (e) True. The 95% CI lies entirely above 50%.

With a random sample, independence is satisfied and the success-failure condition holds. The margin of error is $ME = z^\star\sqrt{\hat{p}(1-\hat{p})/n} = 1.96\sqrt{(0.56)(0.44)/600} = 0.0397 \approx 4\%$.

(a) No. The sample is limited to students who took the SAT *and* who completed an optional web survey, so it does not represent all high-school seniors. (b) $(0.5289, 0.5711)$. We are 90% confident that between 53% and 57% of SAT-taking high-school seniors are fairly certain they will study abroad. (c) In repeated random sampling, about 90% of such 90% CIs would capture the true proportion. (d) Yes — the interval lies entirely above 50%.

(a) We test $H_0 : p = 0.5$ vs. $H_A : p \neq 0.5$, with $\hat{p} = 0.55$ and $n = 617$. Independence holds (random sample); the success-failure condition holds using $p_0 = 0.5$ (both $np_0 = 308.5$ are well above 10). $SE = \sqrt{p_0(1-p_0)/n} \approx 0.02$, so $Z = (0.55 - 0.5)/0.02 = 2.5$. Two-tailed p-value $= 2(0.0062) = 0.0124 < 0.05$, so we reject H_0. Because $\hat{p} > 0.5$, we have strong evidence that a majority of Independents support the plan. (b) Generally a hypothesis test and confidence interval agree, so we would expect a 95% CI entirely above 0.5. If the CI's confidence level does *not* match $1 - \alpha$ (e.g., a 99% CI paired with an $\alpha = 0.05$ test), the CI can include 0.5 while the test still rejects.

(a) The sample represents chips manufactured *during that week*; we should be cautious about generalizing to other weeks, since defect rates can drift over time. (b) The parameter is the true fraction of chips produced that week with severe defects. (c) $\hat{p} = 27/212 = 0.127$. (d) The *standard error* (SE). (e) $SE \approx \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{(0.127)(0.873)/212} \approx 0.023$. (f) $\hat{p} = 0.10$ is only about one SE away from 0.127, which is not surprising; the engineer should not be alarmed. (g) Using $p = 0.10$ instead of $\hat{p}$: $SE = \sqrt{(0.10)(0.90)/212} \approx 0.021$ — almost the same, as is typical when the two proportions being plugged in are similar.

(a) The visitors are a simple random sample, so independence holds. Both 64 and $752 - 64 = 688$ exceed 10, so the success-failure condition holds. A normal-based CI is appropriate. (b) $\hat{p} = 64/752 = 0.085$; $SE = \sqrt{(0.085)(0.915)/752} \approx 0.010$. (c) Using $z^\star = 1.65$: $0.085 \pm 1.65(0.010) \to (0.0685, 0.1015)$. We are 90% confident that between 6.85% and 10.15% of first-time visitors will register under the new design.

(a) The parameter is $p_{\text{Asian-Indian}} - p_{\text{Chinese}}$. The statistic is $\hat{p}_{\text{Asian-Indian}} - \hat{p}_{\text{Chinese}} = 223/4373 - 279/4736 \approx -0.008$. (b) The standard error of the difference is roughly 0.005. (c) $H_0 : p_{\text{Asian-Indian}} - p_{\text{Chinese}} = 0$ vs. $H_A : \neq 0$. The evidence is borderline; there is not strong evidence of a difference in the current-smoker rate between these two ethnic groups.

(a) The standard error of $\hat{p}_{\text{Filipino}} - \hat{p}_{\text{Chinese}}$ is roughly 0.00625. (b) We are 95% confident that the true proportion of Filipino-American current smokers is between 5.28 and 7.72 percentage points higher than that of Chinese Americans. (c) With small changes in rounding, the interval is essentially the same: 5.2 to 7.7 percentage points higher.

(a) The two graphs have similar standard errors (about 0.012), but different centers: Method A is centered at 0.07 (the observed difference in sample proportions), while Method B is centered at 0. Method A is a *bootstrap* distribution (used for a CI), and Method B is a *null* distribution (used for a hypothesis test). (b) A confidence-interval question: what is the difference in the proportions of Bachelor's vs. Associate's students who believe COVID-19 will negatively affect their degree completion? (c) A hypothesis-test question: does the proportion of Bachelor's students who believe COVID-19 will negatively affect degree completion differ from the proportion of Associate's students?

(a) Convert to counts: Nevirapine has 26 Yes / 94 No; Lopinavir has 10 Yes / 110 No. (b) $H_0 : p_N = p_L$ (no difference in virologic-failure rates); $H_A : p_N \neq p_L$. (c) Random assignment gives independence within groups; the study population may not generalize beyond similar patients. The pooled proportion is $\hat{p}_{\text{pool}} = 36/240 = 0.15$, and the success-failure condition is met using this pooled estimate. $Z \approx 2.89$ gives p-value $\approx 0.0039 < 0.05$, so we reject H_0: there is strong evidence of a difference in failure rates between the two treatments.

(a) $SE = \sqrt{(0.79)(0.21)/347 + (0.55)(0.45)/617} \approx 0.03$. Using $z^\star = 1.96$: $(0.79 - 0.55) \pm 1.96(0.03) \to (0.181, 0.299)$. We are 95% confident that the proportion of Democrats supporting the plan is between 18.1 and 29.9 percentage points higher than the proportion of Independents. (b) True.

(a) We treat each job as a Bernoulli trial for "men paid more than women", with $H_0 : p = 0.5$ and $H_A : p \neq 0.5$. (b) Independence is not strictly checkable since the 21 jobs are not a random sample, but the jobs seem to be distinct enough that independence is not unreasonable. The success-failure condition uses $p_0 = 0.5$: both $np_0 = 21(0.5) = 10.5$ exceed 10. Compute $\hat{p} = 19/21 = 0.905$ and $SE = \sqrt{(0.5)(0.5)/21} \approx 0.109$, giving $Z = (0.905 - 0.5)/0.109 \approx 3.72$. The two-tailed p-value is about $2(0.0001) = 0.0002 < 0.05$, so we reject H_0: men are paid more in a discernibly higher fraction of jobs than would be expected by chance.

Conditions fail. Only 4 of 50 in the treatment group and 3 of 25 in the control group yawned — far short of 10 successes and 10 failures in *each* group. Therefore the sampling distribution of $\hat{p}_T - \hat{p}_C$ is not approximately normal, and a normal-based confidence interval is not appropriate. A simulation-based (randomization or bootstrap) interval would be the honest choice here.

(a) False. The confidence interval includes 0, so we cannot conclude the two proportions differ. (b) False. The correct interpretation is that we are 95% confident that Americans making less than $40,000 are between 16 percentage points less and 2 percentage points more likely to be "not at all personally affected" than those making $40,000 or more. (c) False. A *lower* confidence level produces a *narrower* interval. (d) True.

(a) Type I. (b) Type II. (c) Type II.

No — the two samples are not independent. The beginning-of-semester and end-of-semester surveys are taken on the *same* students, so the responses are paired.

(a) With a normal centered at -0.1 and $SD = 0.15$, $P(Z < -2 \times SE) = P(Z < -0.30$ on original scale$) \approx 0.09$. (b) With center -0.4 and $SD = 0.145$, the area below $-2 \times SE = -0.29$ is ≈ 0.77. (c) With center -0.1 and $SD = 0.0671$, the area below $-2 \times SE$ is ≈ 0.07. (d) With center -0.4 and $SD = 0.0678$, the area below $-2 \times SE$ is ≈ 1.00. (e) The larger the true effect δ and the larger the sample size (which shrinks the SE), the more likely a future study is to reject the null hypothesis — i.e., the higher the power.

StatLens Exercises

(a) The parameter is p, the true proportion of *all* students at this campus who skip breakfast. ($\hat{p} = 99/150 = 0.66$ is the sample estimate.) (b) $H_0 : p = 0.60$ vs. $H_A : p > 0.60$ — **one-sided**, because the claim gives a direction ("more than 60%") stated before the data were collected. (c) Using $p_0 = 0.60$: $np_0 = 150(0.60) = 90$ and $n(1 - p_0) = 150(0.40) = 60$. Both are far above 10, so the success-failure condition is met and the normal-approximation z-procedure is trustworthy.

(a) $H_0 : p = 0.10$ vs. $H_A : p < 0.10$ — the claim is a *lower* complication rate. (b) With $n = 62$ and $p_0 = 0.10$: $np_0 = 6.2$, which is **below 10**. The success-failure condition fails. (c) The normal-approximation z-test is not trustworthy here. This is exactly why we used a *bootstrap* interval for this same dataset earlier: when the counts are too small for the normal model, simulation is the honest tool. Sometimes the correct answer to "run the z-test" is "the conditions don't support it."

(a) For the stent data, the two 95% intervals are very close — typically within about a percentage point at each endpoint. With $n = 224$ and many events in each arm, the normal model fits well and both methods agree. (b) For the medical-consultant data the intervals are noticeably *different*, and the bootstrap distribution is visibly right-skewed (the sample proportion is small, so resamples pile up near the low end). A symmetric normal-based interval cannot capture that skew. (c) The normal approximation works when $n\hat{p} \geq 10$ *and* $n(1 - \hat{p}) \geq 10$. The stent data clears this easily, so the sampling distribution of $\hat{p}$ is nearly normal and the formula matches the simulation. The medical-consultant data has only a handful of complications ($n\hat{p} < 10$), so the sampling distribution is skewed — the formula's symmetric interval is the wrong shape, while the bootstrap follows the actual (skewed) distribution. (d) The two methods aren't disagreeing on a fact; they are two *approximations*, and they converge as conditions improve. A disagreement is a **signal that the normal conditions have failed**, not that someone made an error. When they diverge, trust the **simulation**: the bootstrap makes no normality assumption, so it stays honest exactly where the formula breaks down. This is the same reason the course teaches simulation *before* the formulas — simulation is the more general tool; the z-procedure is a convenient shortcut that applies only when conditions are met.

E.13 Chapter 13 — Confidence Intervals for Means

Solutions to the odd-numbered exercises in Chapter 13 — Confidence Intervals for Means.

(a) Statistic: the average nightly sleep among the 25 New Yorkers sampled. Parameter: the average nightly sleep among all New Yorkers. (b) Statistic: the average height of the undergraduates in the sample. Parameter: the average height of all undergraduates at the two universities (or, more broadly, the population the samples represent).

(a) Use the sample mean to estimate the population mean: 171.1 cm. Use the sample median to estimate the population median: 170.3 cm. (b) Use the sample standard deviation, 9.4 cm, and the sample IQR, $177.8 - 163.8 = 14$ cm. (c) $z_{180} = (180 - 171.1)/9.4 \approx 0.95$ and $z_{155} \approx -1.71$. Neither is more than two standard deviations from the mean, so neither height is unusual. (d) No. Sample statistics estimate the population parameter, but they vary from one sample to the next; a new sample would give slightly different values. (e) The **standard error of the mean** measures how much the sample mean varies from sample to sample. Using the given sample, $SE_{\bar{x}} = s/\sqrt{n} = 9.4/\sqrt{507} \approx 0.42$ cm.

(a) The kindergartners will have a **smaller** standard deviation of heights — a group of same-age children is more homogeneous in height than a mixed-age group of adults. (b) The standard error depends on the individual-level variability. For adults, $SE_{\bar{x}} \approx 9.4/\sqrt{100} = 0.94$ cm. For kindergartners, the SE will be **smaller** because their individual heights vary less to begin with.

(a) $df = 5$, $t^\star_5 = 2.02$. (b) $df = 20$, $t^\star_{20} = 2.53$. (c) $df = 28$, $t^\star_{28} = 2.05$. (d) $df = 11$, $t^\star_{11} = 3.11$.

(a) SE ≈ 0.1 weeks (roughly the standard deviation of the bootstrap distribution of $\bar{x}$). (b) A 99% percentile bootstrap CI is roughly (38.45 weeks, 38.85 weeks) — the middle 99% of the bootstrap means. (c) A 99% SE bootstrap CI is $\bar{x} \pm 2.58 \cdot SE \approx (38.49, 38.91)$ weeks. Both intervals give essentially the same story about the true mean gestation length.

The interval is symmetric about the sample mean, so $\bar{x}$ = midpoint = $(18.985 + 21.015)/2 = 20$. The margin of error is $ME = 21.015 - 20 = 1.015$. With $df = 36 - 1 = 35$ and 95% confidence, $t^\star_{35} \approx 2.03$. Using $ME = t^\star \cdot s/\sqrt{n}$, solve for s: $s = ME \cdot \sqrt{n}/t^\star = 1.015 \cdot 6/2.03 \approx 3.0$.

Because $z^\star$ is **smaller** than $t^\star_{df}$, using $z^\star$ would produce a **narrower** interval. A narrower interval captures the true mean less often in repeated sampling, so the actual coverage probability would be **lower than the stated confidence level** (a nominal 95% interval built with $z^\star$ actually covers less than 95% of the time). The discrepancy is largest for small n and shrinks as n grows and $t^\star$ approaches $z^\star$.

We use a **hypothesis test** (two-sample t-test) to evaluate whether the data provide convincing evidence of a difference between two population means, and we use a **confidence interval** to estimate this difference.

(a) A 90% percentile bootstrap CI takes the 5th and 95th percentiles of the bootstrap distribution — read off the histogram, roughly (0.1 m/sec, 0.35 m/sec). (b) A 90% SE bootstrap CI is $\bar{x}_{\text{diff}} \pm 1.65 \cdot SE_{\text{boot}}$, where $\bar{x}_{\text{diff}}$ is the observed difference in means and SE_{boot} is the standard deviation of the bootstrap distribution. Both intervals should be similar because the bootstrap distribution is roughly symmetric. Both suggest Western fence lizards run faster on average than Sagebrush lizards, since the intervals lie entirely above 0.

Using the observed difference in sample means and the SE reported on the bootstrap distribution ($SE_{\text{boot}} = 4.64$), together with $df = \min(22, 22) = 22$ and $t^\star_{22} \approx 2.07$, the 95% mathematical CI is approximately $-12.03 \pm 2.07 \cdot 4.64 \rightarrow$ (\$-2.96, \$-22.42\$ per carat) — that is, we are 95% confident the population average price per carat of 0.99-carat diamonds is \$2.96 to \$22.42 **lower** than that of 1-carat diamonds. (This closely matches the bootstrap intervals in Exercise 16.)

(a) $\mu_{\bar{x}_1} = 15$, $\sigma_{\bar{x}_1} = 20/\sqrt{50} \approx 2.83$. (b) $\mu_{\bar{x}_2} = 20$, $\sigma_{\bar{x}_2} = 10/\sqrt{30} \approx 1.83$. (c) $\mu_{\bar{x}_2 - \bar{x}_1} = 20 - 15 = 5$; $\sigma_{\bar{x}_2 - \bar{x}_1} = \sqrt{(20/\sqrt{50})^2 + (10/\sqrt{30})^2} \approx 3.37$. (d) Thinking of $\bar{x}_1$ and $\bar{x}_2$ as independent random variables, the SD of their difference is the square root of the sum of their squared SDs: $SD_{\bar{x}_2 - \bar{x}_1} = \sqrt{SD^2_{\bar{x}_2} + SD^2_{\bar{x}_1}}$.

(a) True — a paired analysis reduces to a one-sample analysis on the differences. (b) True. Paired data require one observation per unit in each group, so the two groups must have the same size. (Same size alone is not enough for pairing, though.) (c) True — that natural one-to-one correspondence is the definition of paired data.

(d) False. In a paired analysis, we compute the difference between each *matched pair* of observations, not the difference between an observation and the other dataset's average.

(a) Paired — Intel's and Southwest's stock prices are recorded on the same 50 days, so each day pairs one Intel price with one Southwest price. (b) Paired — each of the 50 items has a Target price and a Walmart price for that same item. (c) Not paired — two independent random samples of 100 students are drawn from two different high schools, with no natural correspondence between individual students in one sample and individual students in the other.

(a) The bootstrap distribution of the mean difference (read - write) is centered near $\bar{x}_{r-w} = -0.545$. A 95% percentile CI takes the middle 95% of the histogram, giving approximately $(-1.79, 0.69)$ points. (b) The bootstrap SE (standard deviation of the bootstrap distribution) is roughly $s/\sqrt{n} = 8.887/\sqrt{200} \approx 0.63$. A 95% SE CI is $-0.545 \pm 2 \cdot 0.63 \to (-1.81, 0.71)$. (c) We are 95% confident that in the population of students who take the HSB survey, the true average difference between reading and writing scores (read $-$ write) is somewhere between about -1.8 and 0.7 points. (d) No — both intervals contain 0, so the data do not provide convincing evidence of a real difference in average reading and writing scores.

(a) With $n = 200$, $df = 199$ and $t^\star_{199} \approx 1.97$ for 95% confidence. $SE = 8.887/\sqrt{200} \approx 0.628$. CI: $-0.545 \pm 1.97 \cdot 0.628 \to (-1.78, 0.69)$ points. (b) We are 95% confident that the true average difference between reading and writing scores (read $-$ write) in the population is between -1.78 and 0.69 points. (c) No — 0 lies inside the interval, so the data do not provide convincing evidence of a real average difference between the two scores.

With $\bar{x}_{\text{diff}} = 12.5$, $s_{\text{diff}} = 7.2$, and $n = 50$, we have $df = 49$ and $t^\star_{49} \approx 2.68$ for 99% confidence. $SE = 7.2/\sqrt{50} \approx 1.02$. The 99% CI is $12.5 \pm 2.68 \cdot 1.02 \to (9.77, 15.23)$ feet. We are 99% confident that the average growth of these young trees between 2009 and 2019 was between 9.77 and 15.23 feet. Because the interval lies entirely above 0, the data provide strong evidence of positive average growth over the decade.

StatLens Exercises

(a) The parameter is μ, the *population* mean daily caffeine intake (mg) among all undergraduates at the university. The sample mean $\bar{x}$ from the 48 students is the estimate, not the parameter itself. (b) **Two-sided.** A confidence interval is two-sided by construction — it estimates the parameter with a margin of error in both directions. One-sided procedures live on the hypothesis-testing side, for directional claims. (c) We rarely (essentially never) know σ in practice. The t-procedure plugs in s as the best available estimate of σ and widens the interval slightly (using $t^\star$ instead of $z^\star$) to account for the extra uncertainty introduced by that estimate. The smaller the sample, the heavier the t tails and the wider the interval.

(a) The one-sample t-interval requires (i) **independence** — a random sample from the population (or the result of random assignment in an experiment) — so one observation's value says nothing about another's; and (ii) **approximate normality** — either the individual observations look roughly bell-shaped and outlier-free (for small n), or the sample is large enough (typically $n \geq 30$) for the CLT to smooth things out. (b) At $n = 12$ the CLT is not in play, so inspect the *data* directly with a dotplot or histogram. Look for **strong skew**, **gaps**, or **outliers** — any of those would make the t-interval untrustworthy. (c) No. A strong skew or an outlier at $n = 12$ can pull $\bar{x}$ and inflate s, distorting the interval. Switch to a **bootstrap CI for the mean** (or investigate whether the outlier is a data-quality issue before proceeding). When small samples and non-normal data collide, simulation is the honest choice.

(a) Each runner is their own control, so the difference $D_i = \text{after}_i - \text{before}_i$ subtracts out the huge runner-to-runner variability in baseline ability — variability that would otherwise dominate the standard error. Only the within-runner change (typically much smaller) remains. (b) The paired interval uses $s_D/\sqrt{n}$, where s_D is the SD of the *within-runner differences* — typically small. The two-sample interval ignores the pairing and pools SDs across runners — typically much larger. Same numbers, but the two-sample interval is substantially **wider** and may include 0 when the paired interval excludes it. (c) Compare the mean *change* in 5K time of the trained runners to the mean *change* (or, if no baseline was recorded for the untrained group, the mean final time) of the untrained runners — a **two-sample** t. Matching is impossible because the untrained runners are different people; there is no natural correspondence. (d) The two procedures answer **different questions**. Paired analyzes one variable (the within-pair difference); two-sample analyzes two variables (the two group means). Both use the t distribution because both involve a sample mean with an estimated standard error, but the standard errors are

computed from different things — and calling paired "just a special case" is exactly the confusion that leads people to throw away the pairing (and the power that comes with it).

E.14 Chapter 14 — Hypothesis Tests for Means

Solutions to the odd-numbered exercises in Chapter 14 — Hypothesis Tests for Means.

(a) $df = 10$, two-tailed $p = 2 \cdot P(T_{10} \geq 1.91) \approx 0.085$. Do not reject H_0. (b) $df = 16$, two-tailed $p = 2 \cdot P(T_{16} \geq 3.45) \approx 0.003$. Reject H_0. (c) $df = 6$, two-tailed $p = 2 \cdot P(T_6 \geq 0.83) \approx 0.438$. Do not reject H_0. (d) $df = 27$, two-tailed $p = 2 \cdot P(T_{27} \geq 2.13) \approx 0.042$. Reject H_0.

(a) Point estimate of the average pregnancy length is the sample mean; point estimate of the median is the sample median. (b) $H_0 : \mu = 40$ weeks vs. $H_A : \mu \neq 40$ weeks. With $n = 1000$ the sample mean is well below 40 (roughly 38.7 weeks) and $SE = s/\sqrt{n}$ is small (≈ 0.08), so T is many standard errors below 0 and the p-value is essentially 0. Reject H_0: the average length of gestation in this population is not 40 weeks. (c) The data alone cannot distinguish the two friends' claims. Both a genuine recent decrease in gestation length and a difference in how "40 weeks" is defined would produce a sample mean below 40 in this dataset — we would need historical comparison data (to see a trend) or data on measurement conventions to tell the two explanations apart.

(a) $H_0 : \mu = 8$ (New Yorkers sleep 8 hours per night on average) vs. $H_A : \mu \neq 8$ (New Yorkers sleep less or more than 8 hours on average). (b) The sample is random, and the reported min and max suggest no extreme outliers, so conditions are reasonable. $SE = 0.77/\sqrt{25} = 0.154$, $T = (7.73 - 8)/0.154 = -1.75$, $df = 24$. (c) Two-tailed p-value ≈ 0.093. If in fact New Yorkers averaged 8 hours of sleep per night, we would see a random sample of 25 with a mean at least this far from 8 about 9.3% of the time. (d) Since $p > 0.05$, do not reject H_0. The data do not provide strong evidence that New Yorkers sleep more or less than 8 hours per night on average. (e) Yes — since we did not reject H_0, a 90% CI would include 8.

(a) One-sample t-test. $H_0 : \mu = 5$ vs. $H_A : \mu \neq 5$ at $\alpha = 0.05$. Random sample, so observations are independent; assume the distribution of years of piano lessons is approximately normal. $SE = 2.2/\sqrt{20} = 0.492$, $T = (4.6 - 5)/0.492 = -0.81$, $df = 19$. The one-tail area is about 0.21, so the p-value is about 0.42. Since $p > 0.05$, do not reject H_0: the data do not provide sufficient evidence that the average number of years of piano lessons differs from 5. (b) Using $SE = 0.492$ and $t^\star_{19} = 2.09$, the 95% CI is $4.6 \pm 2.09 \times 0.492 = (3.57,\ 5.63)$. We are 95% confident that the average number of years students in this city take piano lessons is between 3.57 and 5.63. (c) They agree: we failed to reject H_0, and the null value $\mu = 5$ lies inside the 95% CI.

The hypotheses should be about the population means, not the sample means. Use μ_{AD} and μ_I (not $\bar{x}_{AD}$ and $\bar{x}_I$). The null should be an equality: $H_0 : \mu_{AD} = \mu_I$. And since the baker wants to know whether there is *any* difference, the alternative should be two-sided: $H_A : \mu_{AD} \neq \mu_I$.

$H_0 : \mu_{WF} = \mu_S$ vs. $H_A : \mu_{WF} \neq \mu_S$, where μ is the true mean top speed of each species. The observed difference $\bar{x}_{WF} - \bar{x}_S = 0.7$ m/sec sits far in the right tail of the 1,000 randomized differences (which are centered near 0 and rarely reach ± 0.5). Very few permutations produce a difference of 0.7 or more extreme, so the p-value is very small (well under 0.05). Reject H_0: the data provide strong evidence that the two species differ in average top speed, with the Western fence lizard faster than the Sagebrush lizard on average.

$H_0 : \mu_{0.99} = \mu_1$ vs. $H_A : \mu_{0.99} \neq \mu_1$. **Independence:** both samples are random and much smaller than 10% of their respective populations, so within-sample and between-sample independence are reasonable. **Normality:** the box plots are not extremely skewed and show no extreme outliers, so the sampling distribution of $\bar{x}_{0.99} - \bar{x}_1$ is approximately normal. $T_{22} \approx -2.7$, two-tailed p-value ≈ 0.013. Since $p < 0.05$, reject H_0: the data provide convincing evidence that the average price per carat differs between 0.99-carat and 1-carat diamonds (with 1-carat diamonds priced higher per carat).

$H_0 : \mu_T = \mu_C$ vs. $H_A : \mu_T \neq \mu_C$, where μ_T and μ_C are the mean emotional-exhaustion scores in the treatment and control populations. Conditions are given to hold. $SE = \sqrt{\frac{4.44^2}{30} + \frac{8.87^2}{30}} = \sqrt{0.66 + 2.62} = 1.81$. $T = (15.47 - 32.43)/1.81 \approx -9.4$, $df \approx 29$ (Welch), so the p-value is essentially 0. Reject H_0: the data provide overwhelming evidence that mean emotional exhaustion is lower in the MBI treatment group than in the control group. Because this was a randomized experiment, the reduction can be attributed to the mindfulness-based intervention.

$H_0 : \mu_M = \mu_A$ vs. $H_A : \mu_M \neq \mu_A$, where μ_M and μ_A are the mean city fuel efficiencies (MPG) of manual and automatic 2021 cars. Both samples are random with $n = 25$; check the box plots for extreme skew or outliers. Compute $SE = \sqrt{s_M^2/25 + s_A^2/25}$ from the table, then $T = (\bar{x}_M - \bar{x}_A)/SE$ with $df \approx \min(24, 24) = 24$, and the two-tailed p-value. In this sample the manual-transmission cars have higher mean city MPG than the automatics; the resulting test statistic gives $|T| > 2$ and a p-value below 0.05. Reject H_0: the data provide convincing evidence of a difference between the mean city fuel efficiency of manual and automatic 2021 cars.

$H_0 : \mu_M = \mu_A$ vs. $H_A : \mu_M \neq \mu_A$, where μ_M and μ_A are the mean highway fuel efficiencies (MPG) of manual and automatic 2021 cars. The same conditions and formulas apply as in the city-MPG problem: $SE = \sqrt{s_M^2/25 + s_A^2/25}$, $T = (\bar{x}_M - \bar{x}_A)/SE$ with $df \approx 24$. In this sample manual cars again show a higher mean highway MPG than automatic cars, and the corresponding p-value is small enough to reject H_0 at $\alpha = 0.05$. The data provide convincing evidence of a difference between the mean highway fuel efficiency of manual and automatic 2021 cars.

$H_0 : \mu_T = \mu_C$ vs. $H_A : \mu_T \neq \mu_C$, where μ_T and μ_C are the mean number of items recalled in the distracted (treatment) and control populations. $SE = \sqrt{\frac{1.8^2}{22} + \frac{1.8^2}{22}} = \sqrt{0.147 + 0.147} = 0.543$. $T = (4.9 - 6.1)/0.543 \approx -2.21$, $df \approx 21$, two-tailed p-value ≈ 0.038. Since $p < 0.05$, reject H_0: the data provide convincing evidence that distracted eaters recall fewer lunch items on average than controls. Because subjects were randomized to conditions, the reduction can be attributed to the distraction.

(a) Let diff $= \text{temp}_{2022} - \text{temp}_{1950}$. $H_0 : \mu_{\text{diff}} = 0$ vs. $H_A : \mu_{\text{diff}} \neq 0$. (b) The observed mean difference $\bar{x}_{\text{diff}} = 2.53$°F falls well outside the bulk of the 1,000 randomized differences (which are centered at 0 and rarely exceed ± 1.5°F). (c) Almost no randomized differences are as extreme as the observed one, so the p-value is essentially 0. Reject H_0: the data provide strong evidence that the 90th percentile high temperature is different in 2022 than in 1950 (higher in 2022 at the sampled NOAA stations).

(a) Paired: for each station the 2022 measurement corresponds to exactly one 1950 measurement at the same location, so the two years' observations are not independent. Analyze the 26 differences. (b) $H_0 : \mu_{\text{diff}} = 0$ (no change in 90th percentile high temperature at NOAA stations between 1950 and 2022) vs. $H_A : \mu_{\text{diff}} \neq 0$ (there is a change). (c) The stations are described as representative of the lower 48 states, so independence between stations is reasonable. $n = 26$ is close to 30 and the histogram shows no particularly extreme outliers, so the sampling distribution of $\bar{x}_{\text{diff}}$ is approximately normal. (d) $SE = 2.95/\sqrt{26} = 0.579$, $T = (2.53 - 0)/0.579 \approx 4.37$, $df = 25$, two-tailed p-value ≈ 0.0002. (e) Since $p < 0.05$, reject H_0: the data provide strong evidence that the 90th percentile high temperature at NOAA stations is different — specifically, hotter — in 2022 than in 1950. (f) A Type I error is possible: we would have rejected H_0 when in fact there was no true change in average 90th percentile high temperature. (g) No — since we rejected H_0 with a null value of 0, a corresponding confidence interval would lie entirely above 0.

(a) *Paired design:* each student takes both exams; randomly assign, for each student separately, which exam is taken with music and which in silence. Analyze the within-student differences in exam scores. (b) *Independent-samples design:* each student is randomly assigned to use one study environment for *both* exams. Compare the mean exam scores across the two groups of students.

(a) $H_0 : \mu_{\text{diff}} = 0$ vs. $H_A : \mu_{\text{diff}} \neq 0$, where diff $= \text{admissions}_{6\text{th}} - \text{admissions}_{13\text{th}}$. Using the reported sample statistics (with the paired sample sizes and standard deviation of the differences), $T \approx -2.71$, $df = 5$, two-tailed p-value ≈ 0.042. Since $p < 0.05$, reject H_0: the data provide evidence of a difference in average ER admissions between the two Fridays, with admissions higher on Friday the 13th on average. (b) The corresponding 95% CI for μ_{diff} is approximately $(-6.49, -0.17)$ admissions. (c) The reported conclusion overreaches. This is an observational study, so we cannot conclude that Friday the 13th *causes* additional admissions — there may be confounding factors, and the sample size is very small ($n = 6$ pairs). The observed difference is discernible, but we should describe it as an association, not evidence for the "stay home" recommendation.

StatLens Exercises

(a) The parameter is μ, the population mean calorie content of *all* bars produced by this manufacturer. The sample mean $\bar{x}$ from the 36 measured bars is the *estimate* of μ, not the parameter itself. (b) $H_0 : \mu = 220$ vs. $H_A : \mu \neq 220$ — **two-sided**, because the consumer group is asking whether the mean differs from 220, with no direction specified in advance. (c) If the group suspects under-filling, the alternative becomes $H_A : \mu < 220$

(one-sided). The practical danger of switching to a one-sided test *after* looking at the data is that you double the effective Type I error rate: a one-sided test at $\alpha = 0.05$ has a 5% false-alarm rate only when the direction was chosen *before* the data — otherwise both tails are effectively "in play" and you should report the two-sided p-value.

(a) (i) **Independence** of observations (random sample or random assignment). (ii) **Approximate normality of the data** for small n, or of the sampling distribution of $\bar{x}$ (the CLT covers you for larger n, where "larger" depends on how skewed the population is). (b) At $n = 15$ the CLT does little work, so look at the picture itself: roughly symmetric, no heavy tails, no outliers? Skew, gaps, or outliers are warning signs that the t distribution won't accurately model $\bar{x}$. (c) Ask two questions: (i) *Is the outlier a data-quality error?* (typo, sensor failure, wrong unit) — if so, fix or remove it and document why. (ii) *Is the outlier a real observation that simply doesn't fit a normal model?* — in which case the t-test is not the right tool: switch to a **randomization test** (no normality assumption), or report results both with and without the outlier and be transparent about the difference.

(a) For well-behaved, moderate-n data the two p-values are typically very close — often agreeing to two decimal places. The t distribution does a good job of approximating the permutation distribution because the CLT has kicked in and both groups look roughly normal. (b) In a small-n skewed/outlier case the t-test's p-value can be **misleadingly small**: a single outlier pulls one group's mean and inflates the test statistic, while the randomization distribution — built from the actual permutations of the actual data — knows there are not many ways to produce such an extreme split. The randomization p-value is typically *larger* (more conservative) than the t p-value in this regime. (c) The t-test's p-value comes from a t distribution, which is the *exact* sampling distribution of $\bar{x}_1 - \bar{x}_2$ only when both populations are normal. Drop normality and the t distribution becomes an approximation; with small n and skew/outliers, the approximation is poor. The randomization test makes no normality assumption — it asks only that under H_0 the labels are *exchangeable*, an assumption satisfied by random assignment alone. (d) The two procedures aren't "disagreeing about a fact"; they are two procedures with different assumptions. When they diverge, that is a **signal** that the t-test's assumptions have failed. Trust the **randomization test** in that situation: it stays honest exactly where the t distribution breaks down.

E.15 Chapter 15 — Chi-Square Goodness of Fit

Solutions to the odd-numbered exercises in Chapter 15 — Chi-Square Goodness of Fit.

(a) False. The chi-square distribution has a single parameter, the degrees of freedom. (b) True. (c) True — it is a sum of squared standardized deviations, so it cannot be negative. (d) False. As the degrees of freedom increases, the distribution becomes *less* skewed and more symmetric (approaching normal in shape).

(a) H_0: $p_{\text{hard}} = 0.60$, $p_{\text{print}} = 0.25$, $p_{\text{online}} = 0.15$. H_A: at least one of these proportions differs from the claimed value. (b) Expected counts: hard copy $= 126 \times 0.60 = 75.6$, print $= 126 \times 0.25 = 31.5$, online $= 126 \times 0.15 = 18.9$. (c) Independence (a random sample of the professor's students, and each student's format choice is independent of the others) and expected counts ≥ 5 in every category (75.6, 31.5, 18.9 — all satisfied). (d) $\chi^2 = \dfrac{(71-75.6)^2}{75.6} + \dfrac{(30-31.5)^2}{31.5} + \dfrac{(25-18.9)^2}{18.9} \approx 0.28 + 0.07 + 1.97 = 2.32$, with $df = 3 - 1 = 2$, giving p-value ≈ 0.31. (e) Since $p \approx 0.31 > 0.05$, fail to reject H_0. The data do not provide convincing evidence that the professor's predicted distribution of book formats is inaccurate.

StatLens Exercises

(a) A **distribution** across categories: the parameters are the five color proportions $(p_R, p_O, p_Y, p_G, p_B)$, which must sum to 1. (b) H_0: the true distribution matches the claim, $(p_R, p_O, p_Y, p_G, p_B) = (0.30, 0.20, 0.20, 0.15, 0.15)$. H_A: at least one of these proportions differs from the claimed value. (c) The χ^2 statistic sums *squared* deviations between observed and expected counts, so it cannot be negative — any departure from H_0, in any category and in any direction, inflates it. There is no meaningful one-sided GOF because there is no single direction in which the data could disagree with a full distributional claim.

(a) The expected count in every category should be at least 5. We use *expected* rather than observed counts because the test relies on a normal approximation to the sampling variability of the counts under H_0; when the expected counts are too small, that approximation — and the χ^2 reference distribution built from it — becomes unreliable regardless of what the observed counts happen to be. (b) The χ^2 p-value is not trustworthy. Use a

simulation approach instead — StatLens's `simulate/goodness-of-fit/` draws repeated multinomial samples under H_0 to build the actual null distribution of the χ^2 statistic, making no large-sample assumption. The underlying issue is that with small expected counts the test statistic has a *discrete* distribution that does not match the smooth χ^2 curve, especially in the tails. (c) Yes, you can merge a small-expected category with a substantively similar neighbor. **Gained:** larger expected counts, restoring the validity of the χ^2 approximation. **Lost:** the ability to say which of the merged categories was responsible for any departure from H_0. Merging just to inflate counts (without a substantive rationale for the merge) is data-massaging.

(a) The simulated null distribution of χ^2 is right-skewed, supported on $[0, \infty)$, with the observed χ^2 marked as a vertical reference on the dotplot. Its shape approximates a χ^2 distribution with the same degrees of freedom — when the expected counts are large. (b) For well-behaved data the simulation and the approximation agree very closely, often to two decimal places. This confirms that the approximation is doing what it is supposed to. (c) With small expected counts the simulation p-value is the trustworthy one. The χ^2 approximation breaks down precisely in this regime; the simulation makes no large-sample assumption and just builds the actual null distribution of the test statistic. (d) The framing is backwards. The simulation is the **gold standard**, and the formula is an *approximation* to it that happens to be convenient when the conditions are met. Agreement is confirmation that the formula's conditions hold, not redundancy; disagreement is a *signal* that the formula's conditions have failed, and the simulation is the one to trust. This is exactly the same logic as bootstrap-vs-t in the earlier CI chapters.

E.16 Chapter 16 — Chi-Square Test of Independence

Solutions to the odd-numbered exercises in Chapter 16 — Chi-Square Test of Independence.

(a) The two-way table appears below. (b-i) $E_{\text{patch+group, quit}} = \dfrac{(150)(70)}{300} = 35$; the observed count (40) is higher than expected. (b-ii) $E_{\text{only patch, did not quit}} = \dfrac{(150)(230)}{300} = 115$; the observed count (120) is higher than expected.

Treatment	Quit: Yes	Quit: No	Total
Patch + support group	40	110	150
Only patch	30	120	150
Total	70	230	300

(a) Column proportions: Sun ≈ 0.343, Partial ≈ 0.325, Shade ≈ 0.331. (b) Expected counts by habitat (order: sun, partial, shade). Desert: 40.9, 38.7, 39.4. Mountain: 36.7, 34.8, 35.5. Valley: 36.4, 34.5, 35.1. (c) Yes — the observed proportions of the three sunlight categories differ across habitats. (d) A visible difference is not a formal conclusion; we need a χ^2 test to decide whether the differences exceed what shuffling alone would produce.

The observed dataset should produce a **larger** χ^2 statistic than a single random shuffle, because the shuffle destroys any real association between habitat and sunlight preference while the observed data retains it.

(a) H_0: habitat and sunlight preference are independent. H_A: habitat and sunlight preference are associated. (b) After 1000 shuffles the randomization χ^2 values range from 0 up to roughly 15, and the observed χ^2 sits far to the right of that bulk. (c) The p-value is essentially 0, so we reject H_0. There is strong evidence that habitat and sunlight preference are associated — knowing the habitat tells us something about how lizards choose sun, partial, or shade.

(a) H_0: habitat and sunlight preference are independent. H_A: they are associated. (b) With the larger sample the randomization χ^2 values range roughly from 0 to 25, and the observed statistic again falls far above the null bulk. (c) The p-value is approximately 0; reject H_0. There is convincing evidence of an association between habitat and sunlight preference. (d) Larger samples increase **power** — the probability of detecting a real association when one exists — so a real effect that was borderline with the smaller sample now shows up decisively.

(a) False. The χ^2 distribution has one parameter, degrees of freedom. (b) True. (c) True. (d) False. As df increases, the χ^2 distribution becomes **more** symmetric (and more bell-shaped), not less.

H_0: sleep-quality levels and profession are independent. H_A: they are associated. Observations are independent and the expected counts are all large enough, so a χ^2 test is appropriate. The test statistic is $\chi^2 = 1$ on 2 df, giving p-value ≈ 0.6. Since the p-value is well above $\alpha = 0.05$, we fail to reject H_0: the data do not provide convincing evidence of an association between sleep levels and profession.

(a) H_0: age of Los Angeles residents is independent of shipping-carrier preference. H_A: age and shipping-carrier preference are associated. (b) The conditions are **not** satisfied — several expected counts fall below 5, so the χ^2 approximation is not trustworthy. A randomization χ^2 test (or collapsing sparse categories) would be more defensible.

StatLens Exercises

(a) The two categorical variables are **major** (4 levels) and **caffeine source** (4 levels); the contingency table is 4×4 with 16 cells. (b) H_0: major and caffeine source are independent (knowing one tells you nothing about the other). H_A: major and caffeine source are associated. (c) A test of **independence** applies when a single random sample is drawn and both variables are measured; a test of **homogeneity** applies when separate random samples are drawn from each level of one variable and the second variable is measured within each sample. The arithmetic is identical but the language differs: "are these characteristics related?" vs. "do these subpopulations have the same distribution?" Sampling design dictates which framing to use.

(a) Expected counts should all be at least 5. Expected counts (not observed) are the criterion because the χ^2 approximation is a large-sample argument that depends on the *predicted* cell sizes under H_0. (b) The expected count in cell (i, j) is $\dfrac{(\text{row } i \text{ total}) \times (\text{col } j \text{ total})}{\text{grand total}}$. In words: if the variables were independent, the joint frequency would be the product of the marginal proportions; multiplying by the grand total scales that back to a count. (c) A smallest expected count of 4 is below the guideline, so trust in the χ^2 approximation is shaky. Two defensible alternatives: (i) use the **randomization χ^2 test** (StatLens `simulate/randomization-chisq/`), which builds the null distribution directly without a large-sample assumption; (ii) **collapse categories** that are substantively similar to boost expected counts, accepting the loss of granularity.

(a) For well-behaved data the simulation p-value and the χ^2-distribution p-value agree closely — often to two or three decimal places. (b) With a smaller table or a small cell the two diverge; the simulation is the trustworthy one because the χ^2 approximation can produce artificially small p-values in sparse tables, leading to over-rejection of H_0. (c) In a 2×2 table, the χ^2 statistic equals z^2 from the corresponding two-proportion test; both procedures ask the same question ("is the success rate the same in the two groups?") and give the same two-sided p-value, so the equality is algebraic, not coincidental. (d) The omnibus χ^2 test answers "is there **any** association?" but not "**which** row-column combinations are responsible." Useful follow-ups: examine the standardized residuals $(O - E)/\sqrt{E}$ cell-by-cell (large absolute values flag unusual cells), or run targeted 2×2 sub-tables on the comparisons of interest — with a Bonferroni-style adjustment to α if many are tested.

E.17 Chapter 17 — Analysis of Variance

Solutions to the odd-numbered exercises in Chapter 17 — Analysis of Variance.

Alternative. Large differences between group means inflate the between-group variability (MSG), pushing the F-statistic above 1 and providing evidence against H_0: all population means equal.

(a) The means across the original data are **more variable** than across the randomized data — once host species is scrambled, the group means collapse toward a common center. (b) The within-species standard deviations are about the **same** in both plots; permuting labels doesn't change how spread each species' egg lengths are, only how the group means line up. (c) The **original data** produces the larger F-statistic: bigger between-group variability (MSG) with roughly the same within-group variability (MSE) means a larger MSG/MSE ratio.

H_0: $\mu_1 = \mu_2 = \cdots = \mu_6$ (mean chick weight is the same across all six feed supplements). H_A: at least one pair of feed-type mean weights differs. **Conditions.** Independence: chicks were randomly assigned to feed types and (presumably) housed separately, so independence within and between groups is reasonable. Approximate normality: the within-group boxplots look fairly symmetric with no severe outliers. Constant variance: the group SDs are of comparable magnitude; small differences are plausible from sampling variability given the

small n per group. **Test.** $F_{5,\,65} = 15.36$ with p-value ≈ 0. Reject H_0. The data provide convincing evidence that mean weight differs across at least some of the six feed supplements.

(a) H_0: mean weekly MET is the same across all five coffee-consumption levels. H_A: at least one pair of coffee-level mean METs differs. (b) Independence: no collection detail is given, so we must **assume** the participants are independent of one another (in practice, we'd want to verify). Normality: MET is bounded below by 0 and the group SDs exceed the means, indicating strong right skew — but sample sizes are enormous (n in the thousands per group), so the CLT covers the mean. Constant variance: SDs across the five groups (21.1, 25.5, 22.5, 22.0, 22.0) are similar enough to accept this condition. (c) $F_{4,\,50734} = 5.2$, p-value $= 0.0003$. Reject H_0. The data provide convincing evidence that mean MET differs between at least one pair of coffee-consumption groups.

(a) H_0: mean GPA is the same for all three major groups. H_A: at least one pair of mean GPAs differs. (b) The p-value exceeds 0.05, so we **fail to reject** H_0. The data do not provide convincing evidence that the mean GPA differs across the three major groups. (c) Residual df + group df = $195 + 2 = 197$, so total df $= 197$, and the sample size is $197 + 1 = 198$ students.

(a) The **left** randomization distribution corresponds to **Dataset B**. Dataset B has large within-group spread (SD ≈ 20), so the observed F is small and sits inside the bulk of the null distribution — the red line lands where the shuffled F-values commonly fall. (b) The **right** randomization distribution corresponds to **Dataset A**. Dataset A has small within-group spread (SD ≈ 5), which inflates the observed F; the red line sits far out in the right tail relative to the shuffled null.

(a) The significant F-test tells you that **at least one pair** of population means differs; it does **not** tell you *which* pair(s) or *by how much*. Follow-up pairwise comparisons are needed to see the structure. (b) **Below 95%.** Each individual CI misses with probability ≈ 0.05; with six CIs the chance that *at least one* misses is much higher. If the errors were independent, the joint coverage would be roughly $(0.95)^6 \approx 0.74$ — a joint miss rate near 26%. (c) **Bonferroni** (use α/k per CI), **Tukey HSD** (built for all-pairs after ANOVA), or **Scheffé** (any linear contrasts). Covered in Ch 19. (d) In practice you act on the *full set* of CIs, so the trustworthiness of the *set* — not just each individual CI — is what matters. Six CIs each at 95% deliver a much weaker collective guarantee than the individual number suggests.

StatLens Exercises

(a) Explanatory: fertilizer brand (categorical, **4 levels**). Response: yield (quantitative, bushels/acre). (b) H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4$ — all four population mean yields are equal. H_A : **at least one** μ_i differs from the others (*not* "all four differ from each other"). (c) Comparing k means simultaneously reduces to asking whether the variability **between** group means is larger than the variability **within** groups. The F-statistic = (between-group variance) / (within-group variance). Under H_0 both estimates measure the same noise, so $F \approx 1$. Under H_A real group differences inflate the numerator, pushing F above 1 — so a right-tail F-test on the *variance ratio* answers the question about *means*.

(a) Name the three conditions: **(i) independence** within each group (random sample or random assignment), **(ii) approximate normality** of the response in each group (small n) or via the CLT (larger n), and **(iii) equal variances** across groups (the within-group SDs are roughly the same). (b) Boxes with wildly different **heights** (IQRs) are the tell for a failed equal-variance condition. A common rule of thumb: if the largest group SD is more than $\approx 2\times$ the smallest, the equal-variance assumption is a stretch. (c) With $n_1 = 8$ and a strong right skew, group 1's sample mean is unlikely to be near-normal (the CLT hasn't rescued it yet), so the F-test's p-value can be misleading. A defensible alternative is the **randomization ANOVA** (`simulate/randomization-anova/`): shuffle group labels many times to build the actual null distribution of F — no normality assumption, only exchangeability under H_0.

(a) The simulated null distribution of F is **right-skewed**, supported on $[0, \infty)$, and peaks near 1 (under H_0, MSG and MSE both estimate the same noise, so their ratio ≈ 1). With enough shuffles it tracks the theoretical F curve — provided the ANOVA conditions hold. (b) For well-behaved data the two p-values agree closely. (c) With $n = 6$ and a clear outlier the two p-values can diverge, and the **simulation** is the safer report: the F-distribution p-value can be misleadingly small when normality or equal-variance fails, but the randomization p-value only relies on exchangeability under H_0. (d) Two reasons: (i) understanding the F-distribution — why it has two df parameters, why $F \approx 1$ under H_0 — builds intuition that shuffling alone won't; (ii) once conditions

are known to hold, the F-formula returns a p-value instantly, while the simulation needs thousands of shuffles. On well-behaved data the formula is the practical choice; the simulation is the safety net.

E.18 Chapter 18 — Multiple Comparisons

Solutions to the odd-numbered exercises in Chapter 18 — Multiple Comparisons.

(a) **False.** As the number of groups increases, so does the number of pairwise comparisons, and hence the Bonferroni-modified discernibility level $\alpha^\star = \alpha/K$ *decreases* (not increases). (b) **True.** The residual degrees of freedom equal $n-k$, so they grow as n grows for fixed k. (c) **True.** ANOVA is fairly robust to modest departures from equal variance when the group sample sizes are balanced. (d) **False.** Independence is required regardless of the total sample size — a large n does not fix dependent observations.

StatLens Exercises

(a) The **per-test** rate is the false-positive probability of a single test run in isolation — 0.05 by convention. The **family-wise** rate is the probability of *at least one* false positive across the whole family of tests. Per-test α stays fixed; family-wise α grows with the number of tests. (b) Under independence and all-true nulls, $P(\text{at least one false positive}) = 1 - (0.95)^{10} \approx 0.401$ — about a 40% chance of a spurious "discovery" somewhere in the family, even though each individual test looks safe. (c) The decision being reported is "conditions A and C differ" — a *specific* pair singled out from many. If the family-wise error rate is unbounded, the study cannot credibly claim to have located "the" difference. Uncorrected pairwise tests are how false discoveries slip into multi-group studies.

(a) **Tukey HSD** — the standard follow-up when the omnibus F rejects and you want *all* pairwise comparisons on the same footing; usually more powerful than Bonferroni for this exact setting. (b) **No post-hoc needed** — without an omnibus rejection, doing pairwise tests inflates Type I error without a family-level rejection to justify them. Report the null result. (c) **Bonferroni** across the 2 pre-specified comparisons. Bonferroni is more flexible than Tukey (works for any pre-specified collection), and with only 2 tests it costs little power. (d) **Not a multiple-comparisons problem** — two groups means one test (two-sample t), no correction needed. (e) **Bonferroni** (or an FDR procedure like Benjamini–Hochberg). With 100 tests the family-wise error rate at uncorrected $\alpha = 0.05$ is enormous; genomics-scale problems typically use FDR control, but the principle is the same — correct for multiplicity.

(a) Tukey gives narrower CIs on most datasets — it is designed specifically for all-pairwise comparisons of means and exploits the correlation structure of those comparisons. Bonferroni ignores that structure and simply penalizes by the number of tests. (b) With few pairs (e.g., 3 groups $\to$ 3 pairs) the two methods are closer; Tukey still edges out, but by less. With many pairs (6 groups $\to$ 15 pairs) Tukey pulls further ahead. (c) Bonferroni controls the family-wise error rate for *any* collection of tests, treating each as if independent, so the correction is generic (α/m). Tukey HSD is built specifically for pairwise means from the same ANOVA; it uses the studentized-range distribution, which accounts for the correlation among sample-mean differences ($\bar{x}_1 - \bar{x}_2$ and $\bar{x}_1 - \bar{x}_3$ share $\bar{x}_1$). That extra information yields tighter intervals at the same family-wise coverage. (d) Prefer Bonferroni when: (i) you have a **pre-specified** set of comparisons that are not all pairwise (e.g., treatment vs. control and treatment vs. combination, but not control vs. combination); (ii) the tests mix procedure types (some means, some proportions, some correlations); or (iii) the sample sizes are wildly unbalanced — Tukey's equal-n guarantee weakens in the unbalanced case. (e) Both methods provide honest family-wise control at the same nominal level, but they are not identical — Bonferroni's per-pair threshold is α/m, while Tukey's uses the studentized-range distribution, and the two can cross a per-pair boundary at slightly different points. When they disagree on a particular pair, they are answering slightly different questions — neither is "right." For the standard all-pairwise-of-means question, Tukey is the default; report both if the disagreement matters to the conclusion.

E.19 Chapter 19 — Correlation

Solutions to the odd-numbered exercises in Chapter 19 — Correlation.

(a) Strong relationship, but a straight line would not fit the data. (b) Strong relationship, and a linear fit would be reasonable. (c) Weak relationship, and trying a linear fit would be reasonable. (d) Moderate relationship, but a straight line would not fit the data. (e) Strong relationship, and a linear fit would be reasonable. (f) Weak relationship, and trying a linear fit would be reasonable.

(a) Exam 2. The scatterplot of course grade versus Exam 2 shows less scatter, so Exam 2 is more tightly associated with the course grade. The Exam 1 relationship also looks slightly nonlinear. (b) (Answers may vary.) If Exam 2 is cumulative, it might be a better indicator of overall class performance and therefore a more useful predictor.

(a) $r = -0.7 \to (4)$. (b) $r = 0.45 \to (3)$. (c) $r = 0.06 \to (1)$. (d) $r = 0.92 \to (2)$.

(a) There is a moderate, positive, and linear relationship between shoulder girth and height. (b) Changing units — even for just one variable — does not change the form, direction, or strength of a relationship. Correlation is unit-invariant.

(a) There is a somewhat weak, positive, possibly linear relationship between distance traveled and travel time. Note the clustering near the lower-left corner. (b) Changing units does not change the form, direction, or strength of the relationship: longer distances in miles that go with longer times in minutes will still go together when the units become kilometers and hours. (c) Correlation is unaffected by unit changes, so $r = 0.636$ in either set of units.

In every part, one variable is an exact linear function of the other, so the scatterplot is a perfect line and $r = 1$. (1) carbs $=$ meat $- 3$. (2) carbs $=$ meat $+ 2$. (3) carbs $= 2 \times$ meat. The slope is positive in all three, so the correlation is exactly $+1$ in all three. (Sketch a scatterplot with five mock countries at meat = 10, 20, 50, 75, 100 kg and compute the matching carbs values to see it directly.)

(a) True. Correlation is bounded between -1 and $+1$; strength is measured by $|r|$, so $|-0.90| = 0.90 > 0.5$ indicates a stronger linear relationship. (b) False. Correlation measures the *linear* association between two *numerical* variables. Two categorical variables (or two numerical variables related by a curve) require a different measure of association.

(a) $r = 0.7 \to (1)$. (b) $r = 0.09 \to (4)$. (c) $r = -0.91 \to (2)$. (d) $r = 0.96 \to (3)$.

StatLens Exercises

(a) & (b) Answers will vary by round. A common pattern is that students *over*-estimate $|r|$ on weak correlations (a hint of pattern reads as stronger than it is) and slightly *under*-estimate on very strong ones. (c) r summarizes only the *linear* component of the relationship. Two data sets can share the same r but tell very different stories — a clean line and a curve that both register the same r, for instance, or an outlier that inflates r from near-zero to moderate. A perfectly determined curve like $y = x^2$ on $[-1, 1]$ even has $r = 0$. Always look at the scatterplot before quoting a correlation; the picture catches what the number cannot.

(a) Three warning signs: (i) a **non-linear pattern** (curves, U-shapes) — r measures only the linear piece and will understate a real relationship; (ii) **influential outliers** that visibly move r when included or excluded; (iii) **clusters or sub-groups** within the scatter, which hint that a single r is hiding two different relationships (Simpson's-paradox territory). (b) "There is no *linear* association ($r \approx 0$), but the scatterplot shows a clear curved relationship." Calling this "no association" is misleading — the variables are clearly related; the linear summary simply fails to capture it. (c) **Report both** and discuss the influential observations openly. Removing outliers without mentioning them is data-massaging. Readers should be able to see that r is sensitive here — that sensitivity is itself part of the finding.

(a) A likely confounder is **temperature / summer weather**: hot days drive both ice-cream sales (more buyers) and swimming activity (more opportunity for drownings). Neither variable causes the other; both respond to the same underlying driver. (b) The original r still summarizes a real *observed* association, but it is not a causal one. Once temperature is controlled, ice-cream sales and drownings are no longer correlated — exactly the pattern a common-cause explanation predicts. (c) A **randomized controlled experiment**. Random assignment breaks the confounding link by making the treatment groups similar, on average, on every potential confounder (measured or not). With random assignment, the correlation between treatment and outcome can be read causally. (d) Two reasons: (i) **confounding** — coffee drinkers may differ systematically from non-drinkers on age, income, lifestyle, or baseline health in ways that explain the association; (ii) **reverse causation** — sicker people may avoid

coffee because of their symptoms, not because coffee causes better health. Either alone is enough to undermine a causal reading of an observational finding.

E.20 Chapter 20 — Linear Regression

Solutions to the odd-numbered exercises in Chapter 20 — Linear Regression.

Correlation: no units. Intercept: calories (cal). Slope: cal/cm.

(a) First compute the slope: $b_1 = r \cdot s_y/s_x = 0.636 \times 113/99 \approx 0.726$. Then use the fact that the least-squares line passes through $(\bar{x}, \bar{y})$: $b_0 = \bar{y} - b_1\bar{x} = 129 - 0.726 \times 108 \approx 51$. Line: $\widehat{\text{travel time}} = 51 + 0.726 \times \text{distance}$. (b) Slope: for each additional mile between stops, the model predicts an additional 0.726 minutes of travel time on average. Intercept: a distance of 0 miles corresponds to a predicted travel time of 51 minutes — not meaningful in context (zero distance means no travel); the intercept only anchors the line's height. (c) $R^2 = 0.636^2 \approx 0.40$. About 40% of the variability in travel time is explained by the distance between stops. (d) $\widehat{\text{travel time}} = 51 + 0.726 \times 103 \approx 126$ minutes. (e) Residual $= 168 - 126 = 42$ minutes: the actual travel time is 42 minutes *longer* than the model predicts (the model under-predicts here). (f) No — 500 miles is well outside the range of distances used to fit the model, so predicting there would be extrapolation.

(a) $\widehat{\text{poverty}} = 4.60 + 2.05 \times \text{unemployment\_rate}$. (b) The model predicts a poverty rate of 4.60% for a county with 0% unemployment. No US county has an unemployment rate that low, so this value is not directly meaningful — it just anchors the line. (c) For each 1 percentage-point increase in unemployment rate, the model predicts an average increase of 2.05 percentage points in poverty rate. (d) Unemployment rate explains about 46% of the county-to-county variability in poverty rate. (e) $r = \sqrt{0.46} \approx 0.68$ (positive, since the slope is positive).

(a) The scatterplot shows a negative association, so $r = -\sqrt{R^2/100}$; using the printed R^2, $r \approx -\sqrt{0.72} \approx -0.85$ (negative). (b) $b_1 = r \cdot s_y/s_x \approx -0.85 \times 16.95/26.72 \approx -0.54$; $b_0 = \bar{y} - b_1\bar{x} = 30.88 - (-0.54)(30.83) \approx 47.5$. Line: $\widehat{\text{helmet}} = 47.5 - 0.54 \times \text{lunch}$. (c) Intercept: in a neighborhood where no children receive reduced-fee lunches, the model predicts about 47.5% of bike riders wear a helmet. (d) Slope: for each 1 percentage-point increase in children receiving reduced-fee lunches, the model predicts helmet use to drop by about 0.54 percentage points on average. (e) $\widehat{\text{helmet}} = 47.5 - 0.54(40) \approx 25.9$; residual $= 40 - 25.9 = 14.1$ percentage points. This neighborhood has helmet use about 14 percentage points *higher* than the model predicts for its lunch rate.

(a) $H_0 : \beta_1 = 0$ (father's age and baby's weight are not linearly related). $H_A : \beta_1 \neq 0$. (b) The observed slope 0.005 sits comfortably inside the bulk of the null distribution, so the p-value is *large* (many permuted slopes are at least as extreme). We fail to reject H_0: the data do not provide convincing evidence that father's age is associated with baby's weight, so father's age does not appear to be a useful predictor of baby's weight in these data. (c) Yes — the mathematical-model p-value is also large, and both approaches lead to the same fail-to-reject conclusion.

(a) Predicted weight = $b_0 + b_1 \times 30$ (read b_0 and b_1 from the estimate column of the printed regression table; the result is close to 7.1 lbs, near the mean baby weight). (b) $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$. The reported p-value is large (well above 0.05), so we fail to reject H_0. There is no convincing evidence of a linear relationship between father's age and baby's weight. (c) Because we fail to reject the null of no association, father's age is *not* a useful predictor of baby's weight.

(a) Read the 2.5th and 97.5th percentiles off the bootstrap histogram of slopes. The middle 95% of the bootstrap slopes runs roughly from about -0.003 to about 0.013 lb per year of father's age. (b) With 95% confidence, we estimate that each additional year of father's age is associated with somewhere between a 0.003-lb *decrease* and a 0.013-lb *increase* in mean baby weight. Because 0 is inside the interval, the data are consistent with no linear relationship between father's age and baby's weight.

(a) The bootstrap SE is the standard deviation of the bootstrap slopes; reading the spread of the histogram, $SE \approx 0.004$ lb/year. (b) 95% SE interval: $b_1 \pm 1.96 \cdot SE \approx 0.005 \pm 1.96(0.004) \approx (-0.003,\ 0.013)$. (c) With 95% confidence, we estimate the true slope — the mean change in baby's weight (lbs) per additional year of father's age — lies between about -0.003 and 0.013. Since 0 falls inside the interval, the data do not rule out a slope of zero.

(a) $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$, where β_1 is the population slope for annual murders per million on percent living in poverty. (b) The p-value from the regression output is very small, so we reject H_0 at any conventional level. The data provide convincing evidence that poverty percentage is a useful linear predictor of annual murder rate across metropolitan areas: areas with higher poverty tend to have higher annual murder rates. (c) 95% CI: $b_1 \pm t^*_{18} \cdot SE(b_1) = 2.559 \pm 2.101 \cdot SE(b_1)$; reading $SE(b_1)$ from the output, the interval is roughly $(1.8,\ 3.3)$ murders per million per percentage point of poverty. (d) Yes: the 95% CI does not contain 0, and the hypothesis test rejects H_0; both approaches lead to the same conclusion.

(a) The bootstrap SE is the standard deviation of the 1,000 bootstrap slopes; from the spread of the histogram, $SE \approx 0.35$ murders per million per percentage point of poverty. (b) 90% SE CI: $2.559 \pm 1.645 \cdot 0.35 \approx (1.98,\ 3.14)$. (c) With 90% confidence, we estimate that a one-percentage-point increase in poverty is associated with between about 2.0 and 3.1 additional annual murders per million people on average. Since 0 is not in the interval, poverty percentage is a useful predictor of annual murder rate.

(a) The relationship between cans of beer and BAC is **positive**, roughly **linear**, and **strong** (little scatter around a clear upward trend). (b) Read the coefficients from the regression table: $\widehat{\text{BAC}} = b_0 + b_1 \cdot \text{beers}$, with $b_0 \approx -0.013$ and $b_1 \approx 0.018$. Slope: each additional can of beer is associated with a mean BAC increase of about 0.018 g/dL. Intercept: a person who drank 0 beers is predicted to have a slightly *negative* BAC — physically impossible; the intercept only anchors the fitted line and is not meaningful on its own. (c) $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 > 0$ (or $\neq 0$ for a two-sided test). The p-value in the regression output is very small, so we reject H_0: the data provide convincing evidence that drinking more cans of beer is associated with higher BAC. (d) $R^2 = r^2 \approx 0.89^2 \approx 0.79$: about 79% of the variability in BAC is explained by the number of cans of beer. (e) No — a bar sample would introduce many extra sources of variability the OSU experiment held roughly constant: differences in weight, food consumption, drinking pace, drink strength, time since drinking, and biological variation. Those additional sources of noise would weaken the linear relationship between number of drinks and BAC.

StatLens Exercises

(a) Explanatory: square footage; response: price. The population slope is β_1; the reported 0.15 is b_1, the sample estimate. (b) $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$. (c) "For each additional **square foot** of house, the **mean** price is predicted to increase by about **$150**." (Convert 0.15 thousand-dollars per sqft into $150 per sqft.) The most common student errors are (i) omitting "**mean**" (the slope describes the *average* response, not what any single house will sell for), (ii) omitting **units**, and (iii) reversing cause and effect — the slope describes association, not "adding a square foot causes the price to rise $150."

(a) The four regression conditions are **L**inearity (the true relationship is approximately linear), **I**ndependence (observations are independent), **N**ormality of residuals (residuals are approximately normal), and **E**qual variance of residuals (spread of residuals is constant across x). (b) **Linearity** fails when the residual plot shows a systematic **curved** pattern (a U-shape, arc, or wave) instead of random scatter around 0. **Equal variance** fails when the spread of residuals **grows or shrinks** with x — the classic fan or trumpet shape. (c) No — don't discard automatically. First ask: (i) is it a data-quality error (typo, unit mix-up, instrument failure)? If yes, fix or remove and document the reason. (ii) Is it a real but influential observation the linear model can't accommodate? If yes, the *model* is the problem, not the point — consider a transformation, a non-linear model, or transparently report the with- and without-outlier fits.

(a) For a well-behaved (roughly linear, homoscedastic) dataset the two 95% intervals for β_1 typically agree to within a small fraction of the standard error. (b) On heteroscedastic data (visible fan shape in the residuals), the **bootstrap** interval is typically **wider** than the t-interval, because it correctly accounts for the higher variability at some values of x. The t-interval is *artificially narrow* here — the equal-variance assumption it depends on has failed. (c) The **bootstrap** is more honest: it does not pretend the spread is constant when it isn't. The t-interval looks more precise, but its actual coverage falls below the advertised 95%. (d) Transforming (e.g., log–log) is the right first move when there is a *theoretical* reason to expect linearity on the transformed scale — but sometimes no transformation works, or the transformed slope no longer answers the original question. In those cases the bootstrap CI for the slope on the original scale is the right tool: it gives an honest interval without forcing the data into a shape they don't have.

E.21 Chapter 21 — Prediction and Model Fit

Solutions to the odd-numbered exercises in Chapter 21 — Prediction and Model Fit.

(a) The residual plot will show residuals randomly distributed around 0 with approximately constant variance. (b) The residuals will show a fan shape, with higher variability for smaller x and lower variability for larger x; there will also be several points on the right sitting above the line. The linear model is not a good fit here.

Over-estimate. The residual is $e = y - \hat{y}$, so a negative residual means the observed value is smaller than the predicted value — the model predicted a longer shelf life than the apple actually had.

(a) There is a positive, moderate, roughly linear association between number of calories and amount of protein. Protein is more variable for menu items with higher calorie counts, indicating non-constant variance. Two loose clusters are visible: a small group of low-calorie items in the lower left and a larger group of higher-calorie items on the right. (b) Explanatory (predictor): calories. Response (outcome): amount of protein (in grams). (c) A regression line lets us predict the amount of protein for a menu item given only its calorie count — useful when Starbucks displays calories but not the nutritional breakdown. (d) The residuals fan out with predicted value: items with higher predicted protein show more prediction error than items with lower predicted protein, so the model is more trustworthy at the low end than at the high end.

(a) An outlier in the bottom right. It is far from the center of the data in the x-direction, so it has **high leverage**. It is also **influential**: without that observation, the regression line would have a very different slope. (b) An outlier in the bottom right. It has **high leverage** (extreme x), but it is *not* influential — it sits close enough to where the line would go without it that removing it barely changes the fit. (c) The outlier is near the center of the data in the x-direction, so it has **low leverage**. It is neither high-leverage nor influential; its effect on the slope is minimal.

(a) There is a negative, moderate-to-strong, roughly linear relationship between percent of families who own their home and percent of the population living in urban areas. One clear outlier (the state where 100% of the population is urban) sits far to the lower right; the variability in home-ownership also grows as we move from left to right. (b) The outlier at $(100\%, \text{low})$ is horizontally far from the center of the other points, so it has **high leverage**. It is also **influential**: excluding it would noticeably change the slope of the regression line.

(a) Since $R^2 \approx 52\%$, the correlation is $r = \pm\sqrt{0.52} \approx \pm 0.72$. The sign matches the direction of the association in the scatterplot from an earlier exercise on the same data: taller people tend to be heavier, so r is **positive**, $r \approx +0.72$. (b) The residuals show slightly larger spread above zero than below, so they are not perfectly symmetric — an indication that the normality condition is stressed, though not badly violated. A simple least-squares fit is still reasonable here.

(a) $r = \pm\sqrt{R^2}$; the scatterplot shows a positive association between poverty rate and murder rate, so r is **positive**. Plug in the reported R^2 (roughly 71%) to get $r \approx +\sqrt{0.71} \approx +0.84$. (b) The residual plot shows no obvious curvature, no fan shape, and no extreme outliers, so **L**inearity, **N**ormality, and **E**qual variance all look adequate. **I**ndependence is plausible if the 20 metropolitan areas were sampled independently. A simple least-squares fit is appropriate.

(a) The residual spread grows with the predicted value — small residuals for small predicted values, larger residuals for larger predicted values. This is a fan-shaped pattern, and it flags a violation of the **E**qual-variance (constant variability) condition. (b) Unequal variability does **not** affect the fit of the line itself — the regression line still models the average heart weight at each body weight. But it **does** affect the inference (the p-value and CI for the slope), because the pooled SE used in the t-test underestimates uncertainty in the wide-spread region and overstates it in the narrow region. The reported p-value can be trusted only loosely; if inference is central, a bootstrap CI for the slope is a safer read.

StatLens Exercises

(a) Three features of a good residual plot: (i) random scatter around 0 with no visible trend, (ii) roughly constant spread across x (or across predicted value), (iii) no outliers sitting far from the rest of the cloud. (b) Three bad patterns: **curved pattern** → the linear model is mis-specified; the true relationship is non-linear. **Fan shape** → the equal-variance assumption fails (heteroscedasticity). **One huge residual** → a possible outlier or data-quality

issue worth investigating. (c) Small residuals mean the points sit close to the line, but say nothing about whether the line is the *right shape* of curve. A mis-specified model (e.g., a line fit to a curved relationship) can have small residuals near the middle of the data range and still be badly wrong. (d) The **equal-variance** (homoscedasticity) condition has failed. The reported t-statistic and CI use a single pooled SE, which under-states uncertainty where spread is wide and over-states it where spread is narrow. The t-CI's actual coverage is below 95% — a **bootstrap CI for the slope** is more honest in this situation.

(a) It is not just unreliable, it is conceptually wrong: month 25 lies outside the data range (1–12), and there is no evidence that whatever trend held inside the data continues outside it. Worse, month is periodic (month 13 wraps back to January), so a straight line is the wrong shape of model for the phenomenon in the first place. (b) **Trusted: 5 hours** (interior to the data). **Not trusted: 15 hours** (extrapolation — possible but unsupported). **Nonsensical:** -2 **hours** (extrapolation *and* physically impossible). (c) Extrapolating tells you what the *fitted line would predict* outside the range of x, but it tells you *nothing* about whether the linear relationship still holds there. The model has no evidence for or against any behavior outside the observed data. (d) The model was fit on homes in the 500–3500 sqft range. The relationship between price and area at very large sizes is almost certainly different (different buyer pool, fewer comparables, non-linear scaling). The prediction interval's width is computed *as if* the linear model holds at $x = 50000$ — but that assumption is itself the dominant source of error, and the reported width does not reflect it.

(a) **Leverage** measures how far an observation's x-value is from $\bar{x}$ relative to the spread of x. A point at extreme x sits at the end of the "lever arm" and has the *potential* to pull the slope a lot — leverage is a *position* property, decided by x alone, before y is even considered. (b) **The slope changes very little.** A high-leverage point that sits *on* the fitted line simply confirms the trend; removing it barely moves the fit. High leverage in isolation is not a problem. (c) **The slope changes a lot.** A point with high leverage *and* a large residual is called an **influential** observation: its presence pulls the regression line toward itself, distorting the fit. (d) Before deciding to keep or drop a high-leverage point, ask: (i) **Is it a data-quality error?** (typo, mis-recorded value, unit mismatch) — if so, correct or remove it and document the reason. (ii) **Is it a real but extreme observation from the same population?** — if so, report the analysis both with and without the point, and be transparent about how sensitive the conclusion is to its inclusion.

E.22 Chapter 22 — ANOVA for Regression

Solutions to the odd-numbered exercises in Chapter 22 — ANOVA for Regression.

(a) $df_{\text{Reg}} = 1$, $df_{\text{Res}} = n - 2 = 58$, $df_{\text{Total}} = n - 1 = 59$. $MSR = 442/1 = 442$. $MSE = 128/58 \approx 2.21$. $F = MSR/MSE \approx 200.3$. $SST = SSR + SSE = 442 + 128 = 570$. (b) $R^2 = SSR/SST = 442/570 \approx 0.775$. About 77.5% of the variability in highway mpg is explained by the linear regression on curb weight. (c) $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$. An F-statistic near 200 on $(1, 58)$ df is enormous compared with the "$F \approx 1$" null expectation, so the p-value is effectively 0 — strong evidence that curb weight is linearly associated with highway mpg.

(a) $t^2 = 3.1^2 = 9.61 = F$ (to two decimal places), confirming the identity numerically. (b) In simple linear regression the model-level F-test and the slope t-test are the *same* test written in different notation: $F_{1,\,df} = t^2_{df}$, so they must always produce identical p-values.

(a) $F = 0.34$ is *below* 1, and the null-expected value is near 1. That puts the observed F in the bulk of the null distribution, so the p-value is large — roughly 0.5 or bigger (a formal computation gives $p \approx 0.56$). (b) No evidence in these data of a linear relationship between voltage and bulb lifetime. (c) A non-rejecting F-test does not prove there is *no* relationship. Two alternative explanations: (i) the relationship exists but is **nonlinear** (a curve or threshold effect), which the linear F-test cannot detect; (ii) a real linear effect exists but the sample of $n = 40$ has too little power to reach significance — a larger study might discern it.

StatLens Exercises

(a) Both tests state $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$. (b) They are the same hypothesis because in simple linear regression there is only *one* slope to test. In multiple regression the F-test tests all slopes jointly ($H_0 : \beta_1 = \beta_2 =$

$\cdots = 0$) while each t-test tests a single slope at a time; there the two tests diverge. (c) $t = \pm\sqrt{F} = \pm\sqrt{12.25} = \pm 3.5$. Because $F_{1,\,48} = t_{48}^2$, this t produces exactly the same p-value.

(a) "About **72%** of the variability in the response is explained by the **linear model on this predictor**." (b) "72% accurate" is a classification idea, not a regression one; individual predictions are not each 72% correct. The correct statement is that 72% of the *variability* in y is explained by the fitted line. (c) $R^2 = 0.72$ does **not** tell you: (i) whether the LINE conditions hold — a curved relationship can still produce a high R^2; (ii) whether the model will generalize to new data — R^2 is computed on the fit sample, not held-out data; (iii) whether the fit is *practically* useful — $R^2 = 0.72$ might be excellent for one problem and poor for another, depending on how much precision the decision at hand requires.

(a) This is the classic **multicollinearity** signature: the predictors are so correlated with each other that no single one is *uniquely* responsible for the fit, but together they explain a discernible share of the variability. The F-test picks up the joint contribution; the individual t-tests each ask "does *this* predictor add something beyond the others?" — and when the predictors are redundant, none does individually. (b) Neither the joint fit nor any single predictor's unique contribution is discernible from the data. Either there is no real relationship, or the sample is too small for the effects that are present to reach significance — more data would help. (c) With correlated predictors and small samples this pattern can arise, but the **model-level F-test is the more trustworthy summary** for whether the model works as a whole; a single significant t against a non-significant F is often a false positive on the individual test. Refit with fewer predictors or collect more data before drawing conclusions.

E.23 Chapter 23 — Multiple Regression

Solutions to the odd-numbered exercises in Chapter 23 — Multiple Regression.

(a) $\widehat{\texttt{interest\_rate}} = 4.31 + 0.041(15) + 0.16(36) + 0.25(2) = 4.31 + 0.615 + 5.76 + 0.50 = 11.19$ (percent). (b) Among borrowers with the same debt-to-income and same loan term, each additional credit check in the last 12 months is associated with an increase of about 0.25 percentage points in the predicted interest rate, holding the other predictors constant. (c) In symbols, $H_0 : \beta_{\text{credit\_checks}} = 0$, given that `debt_to_income` and `term` are already in the model. In words: after accounting for debt-to-income and loan term, the number of recent credit checks has no linear relationship with interest rate. A small p-value means credit checks add explanatory power *beyond* what those two predictors already provide.

(a) Two predictors: **Years of experience** (VIF $= 8.7$) and **Years at company** (VIF $= 9.1$). Both are close to the rule-of-thumb threshold of 10, which flags severe multicollinearity. In context, someone with more overall experience tends also to have more time at the current company — the two variables carry overlapping information. (b) When two predictors move together, the model cannot separate their individual contributions cleanly; the standard error on each coefficient inflates, so the estimated slope can shift wildly with small changes in the data. That makes "the effect of years of experience, holding years at company constant" hard to pin down. (c) Drop one of the two collinear predictors, OR combine them into a single derived variable (for example, total experience or a ratio like years-at-company / years-of-experience). A third option is to keep both but stop interpreting the individual slopes and use the model only for prediction.

(a) No. A p-value of 0.28 is not evidence that the predictor helps; combined with the strong pairwise association with an existing predictor, adding it risks multicollinearity without gaining useful signal. (b) Yes — subject-matter knowledge is a legitimate reason to include a predictor. Statistical non-significance is not evidence of "no effect"; it just means the sample did not resolve one. If theory says the variable matters, keeping it in yields an honest estimate of its coefficient (and its uncertainty) *given* the other predictors. (c) The residual standard error measures how much variability in the response remains unexplained. Essentially unchanged s means the predictor is not reducing residual variability — it is not earning its degree of freedom in a predictive sense.

StatLens Exercises

(a) $\texttt{interest\_rate} = \beta_0 + \beta_1 \cdot \texttt{debt\_to\_income} + \beta_2 \cdot \texttt{term} + \beta_3 \cdot \texttt{credit\_checks} + \varepsilon$. (b) H_0: none of the three predictors is linearly related to interest rate ($\beta_1 = \beta_2 = \beta_3 = 0$). H_A: at least one predictor is linearly related to interest rate. (c) $H_0 : \beta_1 = 0$, given `term` and `credit_checks` are already in the model. This asks a *conditional* question — does `debt_to_income` add anything above and beyond what term and credit checks already explain? The F-test asks the *unconditional* joint question — does any predictor contribute at all?

(a) StatLens reports the fitted equation directly. It will read $\widehat{\text{head\_l}} = b_0 + b_1 \cdot \text{total\_l} + b_2 \cdot \text{skull\_width}$; students should copy the actual numeric coefficients from the tool. (b) Each slope is the mean change in head length per one-unit change in that predictor, **holding the other predictor fixed**. Report the actual numbers and units (mm per mm) from the tool. (c) The multi-predictor total-length slope is smaller in magnitude than the simple-regression slope. Larger possums tend to have both longer total lengths and wider skulls, so the simple slope absorbed some of the skull-width effect. The multi-predictor slope isolates the total-length effect after skull width is accounted for. (d) The model-level F-statistic is large and the p-value is essentially zero — both predictors carry real signal about head length.

(a) The F-statistic is large enough that its p-value falls below 0.05, while every individual t-test has a p-value well above 0.05. This pattern is the fingerprint of multicollinearity: the joint contribution of the predictors is real, but no *single* predictor is uniquely responsible for it. (b) Yes — the pairwise scatterplots show one or more strongly linear pairs. In expert mode, at least two VIFs will be well above 5, confirming what the scatterplots suggested. (c) The model is trustworthy for **prediction**: put in the predictor values, get an interest-rate estimate. It is *not* trustworthy for **interpretation**: statements like "predictor x_1 specifically drives the response" are unreliable because x_1 can be swapped for a linear combination of the other predictors without hurting the fit. (d) Rejecting the model based on individual t-tests would throw out a *jointly informative* model just because its predictors overlap. Better responses are to combine the collinear predictors, drop one of them, or accept the model for prediction and refuse to interpret individual slopes.

E.24 Chapter 24 — Probability Rules

Solutions to the odd-numbered exercises in Chapter 24 — Probability Rules.

(a) False. This is the gambler's fallacy: successive free throws are essentially independent, so her next attempt still has about a 70% chance of going in — past misses do not "compensate." (b) False. Rolling a 4 is a subset of rolling an even number ($\{2, 4, 6\}$); the two events occur together whenever she rolls a 4, so they are not mutually exclusive. (c) True. "It rains" and "it does not rain" are complementary events, and complementary probabilities sum to 1.

In every version, the sample proportion of heads is more concentrated near 0.50 with $n = 100$ than with $n = 10$ (law of large numbers). Choose the flip count that pushes the proportion into the winning region. (a) Larger than 0.60 — **10 flips** (needs a lucky excursion above 0.60; unlikely at $n = 100$). (b) Larger than 0.40 — **100 flips** (with $n = 100$ the proportion is almost always above 0.40). (c) Between 0.40 and 0.60 — **100 flips** (concentration inside the middle interval). (d) Smaller than 0.30 — **10 flips** (again needs a lucky excursion, this time to the low side).

(a) $P(\text{all tails}) = (1/2)^{10} = 1/1024 \approx 0.001$. (b) $P(\text{all heads}) = (1/2)^{10} = 1/1024 \approx 0.001$. (c) $P(\text{at least one tails}) = 1 - P(\text{all heads}) = 1 - 1/1024 = 1023/1024 \approx 0.999$.

Let I = Independent, S = swing voter. $P(I) = 0.35$, $P(S) = 0.23$, $P(I \text{ and } S) = 0.11$. (a) Not disjoint — 11% of voters are both. (b) Venn: only $I = 0.35 - 0.11 = 0.24$; both = 0.11; only $S = 0.23 - 0.11 = 0.12$; neither = $1 - 0.24 - 0.11 - 0.12 = 0.53$. (c) $P(I \text{ only}) = 0.24$, or 24%. (d) $P(I \text{ or } S) = 0.35 + 0.23 - 0.11 = 0.47$, or 47%. (e) $P(\text{neither}) = 1 - 0.47 = 0.53$, or 53%. (f) $P(S \mid I) = 0.11/0.35 \approx 0.314 \neq P(S) = 0.23$, so being a swing voter is **not** independent of being an Independent.

(a) Independent. Two randomly selected students' outcomes are unrelated (assuming independent grading), so one earning an A does not change the probability the other does. Not disjoint — both can earn A's. (b) Neither. Study partners' grades are positively related (shared study time, materials), so their outcomes are not independent. They also can both earn A's, so not disjoint. (c) No. Being able to occur at the same time only rules out disjoint; independence is a separate property (about whether one event's occurrence changes the probability of the other).

(a) Men with at least a Bachelor's: $0.16 + 0.09 = 0.25$. (b) Women with at least a Bachelor's: $0.17 + 0.09 = 0.26$. (c) Assuming the husband's and wife's education levels are **independent**, $P(\text{both} \geq \text{Bachelor's}) = 0.25 \times 0.26 = 0.065$. (d) The independence assumption is not reasonable — people tend to partner with others of similar educational attainment (assortative mating), so the true probability is likely higher than 0.065.

With $P(A) = 0.3$ and $P(B) = 0.7$: (a) No. Without knowing independence or the joint probability, $P(A \text{ and } B)$ is not determined by $P(A)$ and $P(B)$ alone. (b) Assuming independence: (c) $P(A \text{ and } B) = 0.3 \times 0.7 = 0.21$. (d) $P(A \text{ or } B) = 0.3 + 0.7 - 0.21 = 0.79$. (e) $P(A \mid B) = P(A) = 0.3$ (independence means conditioning on B does not change A's probability). (f) If $P(A \text{ and } B) = 0.1$: independence would require $P(A) \times P(B) = 0.21 \neq 0.1$, so A and B are **not** independent. (g) $P(A \mid B) = P(A \text{ and } B)/P(B) = 0.1/0.7 \approx 0.143$.

Let $W =$ believes earth is warming, $L =$ liberal Democrat. From the joint table, $P(W) = 0.60$, $P(L) = 0.20$, $P(W \text{ and } L) = 0.18$. (a) Not mutually exclusive — 18% are both. (b) $P(W \text{ or } L) = 0.60 + 0.20 - 0.18 = 0.62$. (c) $P(W \mid L) = 0.18/0.20 = 0.90$. (d) For conservative Republicans ($P(W \text{ and } CR) = 0.11$, $P(CR) = 0.33$): $P(W \mid CR) = 0.11/0.33 \approx 0.33$. (e) Not independent. $P(W \mid L) = 0.90$ while $P(W \mid CR) \approx 0.33$ and $P(W) = 0.60$ — the belief probability changes sharply with party/ideology. (f) $P(\text{Mod/Lib Rep} \mid \text{not } W) = 0.06/0.34 \approx 0.18$.

Column totals from the table: 248 males, 252 females (500 total). (a) Not mutually exclusive — six women pick Five Guys. (b) $P(\text{In-N-Out} \mid \text{male}) = 162/248 \approx 0.653$. (c) $P(\text{In-N-Out} \mid \text{female}) = 181/252 \approx 0.718$. (d) Assuming the man's and woman's preferences are **independent**, $P(\text{both In-N-Out}) = 0.653 \times 0.718 \approx 0.469$. The independence assumption is questionable — dating partners often share tastes (or influence each other), so preferences may be correlated. (e) $P(\text{Umami or female}) = P(U) + P(F) - P(U \text{ and } F) = 6/500 + 252/500 - 1/500 = 257/500 = 0.514$.

Let $T =$ attends tutoring, $P =$ passes. $P(T) = 0.70$, $P(P \mid T) = 0.90$, $P(P \mid T^c) = 0.72$. (a) $P(T \text{ and } P) = P(P \mid T)P(T) = 0.90 \times 0.70 = 0.63$. (b) By the law of total probability, $P(P) = P(P \mid T)P(T) + P(P \mid T^c)P(T^c) = 0.90(0.70) + 0.72(0.30) = 0.63 + 0.216 = 0.846$. (c) By Bayes' rule, $P(T \mid P) = P(T \text{ and } P)/P(P) = 0.63/0.846 \approx 0.745$.

Let $D =$ has lupus, and $+ =$ tests positive. $P(D) = 0.02$, $P(+ \mid D) = 0.98$, and "74% accurate if the person does not have the disease" means $P(- \mid D^c) = 0.74$, so $P(+ \mid D^c) = 0.26$. Total probability: $P(+) = 0.02(0.98) + 0.98(0.26) = 0.0196 + 0.2548 = 0.2744$. Bayes' rule: $P(D \mid +) = 0.0196/0.2744 \approx 0.071$. Even after a positive test, there is only about a 7% chance the patient has lupus — the disease is so rare that false positives from the (much larger) healthy population overwhelm the true positives. In that sense, "it's never lupus" is a defensible first guess.

With replacement, each draw sees the original 10-marble bag (5 red, 3 blue, 2 orange). (a) $P(\text{first blue}) = 3/10 = 0.30$. (b) With replacement, the second draw is independent of the first: $P(\text{second blue}) = 3/10 = 0.30$. (c) Same reasoning — with replacement, prior draws don't change the composition: $P(\text{second blue}) = 3/10 = 0.30$. (d) $P(\text{two blue}) = (3/10)^2 = 9/100 = 0.09$. (e) Yes, the draws are independent: replacement resets the bag, so what you drew before does not change the probability of what you draw next.

Without replacement from the 10-chip bag (5 red, 3 blue, 2 orange). (a) After removing 1 blue, 9 chips remain with 2 blue: $P(\text{second blue}) = 2/9$. (b) After removing 1 orange, 9 chips remain with 3 blue: $P(\text{second blue}) = 3/9 = 1/3$. (c) $P(\text{two blue}) = (3/10)(2/9) = 6/90 = 1/15 \approx 0.067$. (d) No — the composition of the bag changes after each draw, so the probability of a color on the second draw depends on what was drawn first.

Total students: 24 with 7 jeans, 4 shorts, 8 skirts, and $24 - 7 - 4 - 8 = 5$ leggings. Draw 3 without replacement; want exactly 1 leggings and 2 jeans:

$$P = \frac{\binom{5}{1}\binom{7}{2}}{\binom{24}{3}} = \frac{5 \times 21}{2024} = \frac{105}{2024} \approx 0.052.$$

Equivalently, $3 \times (5/24)(7/23)(6/22) \approx 0.052$ (three orderings for which slot holds the leggings).

Using the joint counts (Yes-coverage row from the table plus the No-coverage row from the source: Excellent 459, Very Good 524, Good 596, Fair 262, Poor 132, No-total 1,973). Grand total $= 17{,}476 + 1{,}973 = 19{,}449$. (a) $P(\text{excellent AND no coverage}) = 459/19{,}449 \approx 0.024$. (b) $P(\text{excellent OR no coverage}) = P(\text{excellent}) + P(\text{no coverage}) - P(\text{both}) = (4{,}198 + 459)/19{,}449 + 1{,}973/19{,}449 - 459/19{,}449 = 6{,}171/19{,}449 \approx 0.317$.

Let $M =$ manufactured on Machine A, and $D_3 =$ all three inspected parts are defective. Prior $P(A) = 0.40$, $P(B) = 0.60$; defect rates $P(\text{defective} \mid A) = 0.02$ and $P(\text{defective} \mid B) = 0.05$. Assuming independence of the three parts from the same machine, $P(D_3 \mid A) = 0.02^3 = 8 \times 10^{-6}$ and $P(D_3 \mid B) = 0.05^3 = 1.25 \times 10^{-4}$.

(a) By Bayes' rule,

$$P(A \mid D_3) = \frac{P(D_3 \mid A)P(A)}{P(D_3 \mid A)P(A) + P(D_3 \mid B)P(B)} = \frac{0.4 \times 8 \times 10^{-6}}{0.4 \times 8 \times 10^{-6} + 0.6 \times 1.25 \times 10^{-4}} \approx 0.041.$$

About 4.1%. (b) Three consecutive defects are much more likely under Machine B (defect rate 5%) than under Machine A (defect rate 2%): the likelihood ratio is about $0.05^3/0.02^3 = 15.6$, which sharply outweighs the 40/60 prior in Machine B's favor.

E.25 Chapter 25 — Random Variables

Solutions to the odd-numbered exercises in Chapter 25 — Random Variables.

(a) If we view \$X = \$ number of smokers in a random sample of 100 students as Binomial$(n = 100, p = 0.13)$, then $E(X) = np = 100 \times 0.13 = 13$ smokers. (b) No. The 27 students waiting outside the gym at 8:55 am on a Saturday are not a random sample of the university — they self-selected into an early-morning exercise activity, so the smoking rate in this group is likely lower than the campus-wide 13%. The binomial expected-value formula requires the trials to be identically distributed random draws, which fails here.

Let W be the payout. With $\binom{12}{3} = 220$ equally likely three-marble subsets, $P(3\,\text{red}) = \binom{5}{3}/220 = 10/220$, $P(3\,\text{blue}) = \binom{4}{3}/220 = 4/220$, and $P(W = 0) = 206/220$.
(a) $E(W) = 40 \cdot \frac{10}{220} + 25 \cdot \frac{4}{220} + 0 \cdot \frac{206}{220} = \frac{500}{220} \approx \$2.27$. $E(W^2) = 40^2 \cdot \frac{10}{220} + 25^2 \cdot \frac{4}{220} = \frac{18{,}500}{220} \approx 84.09$, so $\text{Var}(W) \approx 84.09 - (2.27)^2 \approx 78.93$ and $\text{SD}(W) \approx \$8.88$.
(b) $N = W - 5$ shifts the mean by -5 and leaves the spread unchanged: $E(N) \approx -\$2.73$ and $\text{SD}(N) \approx \$8.88$.
(c) No. The expected net profit is about $-\$2.73$ per play, so in the long run a risk-neutral player loses roughly \$2.73 each time.

The eight bins have counts $29, 32, 21, 25, 12, 15, 5, 4$, totalling $n = 143$ cats.
(a) The two bins below 2.5 kg contain $29 + 32 = 61$ cats, so about $61/143 \approx 0.43$ (43%).
(b) The bin from 2.5 to 2.75 kg contains 21 cats, so about $21/143 \approx 0.15$ (15%).
(c) The three bins from 2.75 to 3.5 kg contain $25 + 12 + 15 = 52$ cats, so about $52/143 \approx 0.36$ (36%).

A valid probability distribution requires (i) every probability in $[0, 1]$ and (ii) probabilities summing to 1.
(b) $0, 0, 1, 0, 0$: sums to 1 and all values are in $[0, 1]$ — **valid** (a degenerate distribution that assigns all probability to one grade).
(c) $0.3, 0.3, 0.3, 0, 0$: sums to $0.9 \neq 1$ — **invalid**.
(d) $0.3, 0.5, 0.2, 0.1, -0.1$: sums to 1, but -0.1 is not a legal probability — **invalid**.
(e) $0.2, 0.4, 0.2, 0.1, 0.1$: sums to 1 and all values are in $[0, 1]$ — **valid**.
(f) $0, -0.1, 1.1, 0, 0$: sums to 1, but $-0.1 < 0$ and $1.1 > 1$ — **invalid**.

E.26 Chapter 26 — Expected Value and Variance

Solutions to the odd-numbered exercises in Chapter 26 — Expected Value and Variance.

Each scenario is equally likely, so $P = 1/3$ for each. The expected return is $E(R) = \frac{1}{3}(18) + \frac{1}{3}(9) + \frac{1}{3}(-12) = \frac{15}{3} = 5\%$.

Let W = winnings on a \$1 bet on red. $W = \$1$ with probability $18/38 = 9/19$ and $W = -\$1$ with probability $20/38 = 10/19$. $E(W) = 1 \cdot \frac{18}{38} + (-1) \cdot \frac{20}{38} = -\frac{2}{38} = -\frac{1}{19} \approx -\$0.053$. For the standard deviation, $E(W^2) = (1)^2 \cdot \frac{18}{38} + (-1)^2 \cdot \frac{20}{38} = 1$, so $\text{Var}(W) = 1 - (-1/19)^2 = 1 - 1/361 = 360/361$ and $\text{SD}(W) = \sqrt{360/361} \approx \$1.00$. On average you lose about 5 cents per \$1 wagered, with a typical bet-to-bet swing of about a dollar.

Let C = price of coffee and M = price of muffin, with $\mu_C = \$1.40$, $\sigma_C = \$0.30$, $\mu_M = \$2.50$, $\sigma_M = \$0.15$, and $C \perp M$. (a) Daily total $D = C + M$. $E(D) = 1.40 + 2.50 = \$3.90$; $\text{Var}(D) = (0.30)^2 + (0.15)^2 = 0.09 + 0.0225 = 0.1125$; $\text{SD}(D) = \sqrt{0.1125} \approx \$0.34$. (b) Weekly total $W = D_1 + D_2 + \cdots + D_7$ (days independent). $E(W) = 7(3.90) = \$27.30$; $\text{Var}(W) = 7(0.1125) = 0.7875$; $\text{SD}(W) = \sqrt{0.7875} \approx \$0.89$.

Note that the *mean* scales by 7 but the *SD* scales by only $\sqrt{7} \approx 2.65$ — the reason averaging (or, equivalently, larger n) shrinks standard error.

(a) By independence, $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) = 25 + 25 = 50$. Since $\bar{X} = (X_1 + X_2)/2$, $\text{Var}(\bar{X}) = \frac{1}{4}\text{Var}(X_1 + X_2) = \frac{1}{4}(50) = 12.5$. (b) For $n = 2$, $\text{Var}(\bar{X}) = \sigma^2/2$.

(a) In general, $\text{Var}(\bar{X}) = \sigma^2/n$ (the sample-mean variance formula). This follows from $\bar{X} = (1/n)\sum X_i$: $\text{Var}(\bar{X}) = (1/n)^2 \sum \text{Var}(X_i) = (1/n^2)(n\sigma^2) = \sigma^2/n$, using independence. (b) With $\sigma^2 = 36$ and $n = 16$, $\text{Var}(\bar{X}) = 36/16 = 2.25$ and $\text{SD}(\bar{X}) = \sqrt{2.25} = 1.5$.

E.27 Chapter 27 — Binomial Distribution

Solutions to the odd-numbered exercises in Chapter 27 — Binomial Distribution.

(a) Yes. The four binomial conditions are met: (i) $n = 10$ trials, (ii) each 18–20 year old either did or did not consume alcohol, (iii) the success probability $p = 0.697$ is constant across a random sample (10 is a tiny fraction of the U.S. 18–20 population, so the 10% rule for independence holds), (iv) trials are independent. (b) $P(X = 6) = \binom{10}{6}(0.697)^6(0.303)^4 \approx 210 \cdot 0.1147 \cdot 0.00843 \approx 0.203$. (c) "Exactly 4 have *not* consumed" is the same event as "exactly 6 have consumed," so the answer is again ≈ 0.203. (d) With $n = 5$, $P(X \le 2) = P(0) + P(1) + P(2) \approx 0.0026 + 0.0294 + 0.135 \approx 0.167$. (e) $P(X \ge 1) = 1 - P(X = 0) = 1 - (0.303)^5 \approx 1 - 0.0026 \approx 0.997$.

(a) $\mu = np = 50(0.70) = 35$ and $\sigma = \sqrt{np(1-p)} = \sqrt{50(0.70)(0.30)} = \sqrt{10.5} \approx 3.24$. (b) Yes. 45 is $(45 - 35)/3.24 \approx 3.1$ standard deviations above the mean — well into the tail. (c) Using the normal approximation with continuity correction, $z = (44.5 - 35)/3.24 \approx 2.93$, so $P(X \ge 45) \approx P(Z \ge 2.93) \approx 0.0017$. The tiny probability confirms the intuition in (b) — seeing 45 or more would be surprising.

With $n = 3$ spins and $p = 1/4$ for the color of interest: (a) $P(\text{at least one red}) = 1 - (3/4)^3 = 1 - 27/64 = 37/64 \approx 0.578$. (b) $P(\text{exactly 2 reds}) = \binom{3}{2}(1/4)^2(3/4) = 9/64 \approx 0.141$. (c) $P(\text{exactly 1 blue}) = \binom{3}{1}(1/4)(3/4)^2 = 27/64 \approx 0.422$. (d) $P(\text{at most 2 greens}) = 1 - P(3 \text{ greens}) = 1 - (1/4)^3 = 63/64 \approx 0.984$.

Let $p = 0.125$ be the probability of green eyes. (a) $P(\text{1st green, 2nd not}) = 0.125 \cdot 0.875 \approx 0.109$. (b) $P(\text{exactly one of two green}) = \binom{2}{1}(0.125)(0.875) \approx 0.219$. (c) $P(\text{exactly 2 of 6 green}) = \binom{6}{2}(0.125)^2(0.875)^4 \approx 15 \cdot 0.0156 \cdot 0.586 \approx 0.137$. (d) $P(\text{at least one of 6 green}) = 1 - (0.875)^6 \approx 1 - 0.449 \approx 0.551$. (e) $P(\text{1st green is the 4th child}) = (0.875)^3(0.125) \approx 0.670 \cdot 0.125 \approx 0.0837$. (f) With $p_{\text{brown}} = 0.75$, $\mu = 6(0.75) = 4.5$ and $\sigma = \sqrt{6(0.75)(0.25)} \approx 1.06$. Only 2 brown-eyed children is roughly 2.4 SD below the mean; $P(X \le 2) \approx 0.038$. Yes, that would be considered unusual.

(a) With 8 distinct songs and no repeats, there are $8! = 40{,}320$ orderings. (b) By symmetry, each song is equally likely to be first, so $P(\text{Track 1 first}) = 1/8 = 0.125$. (c) Fix Track 1 in slot 1 and Track 2 in slot 8; the remaining 6 songs can fill slots 2–7 in 6! ways. So $P = 6!/8! = 1/(8 \cdot 7) = 1/56 \approx 0.0179$.

(a) $\mu = np = 54(0.93) = 50.22$. (b) $\sigma = \sqrt{54(0.93)(0.07)} = \sqrt{3.5154} \approx 1.875$. (c) With continuity correction, $P(X > 50) = P(X \ge 51) \approx P\left(Z \ge \dfrac{50.5 - 50.22}{1.875}\right) = P(Z \ge 0.149) \approx 0.441$. (d) Even though the airline sold only 54 tickets under the assumption that about 7% will not show, there is roughly a 44% chance the flight is overbooked — expect bumped passengers on roughly two flights out of every five. Selling 54 tickets is quite aggressive.

StatLens Exercises

(a) **Yes** — all four conditions met with $n = 50$ and $p = 0.5$. (b) **No** — drawing *without replacement* changes p after each draw, violating the constant-p and independence conditions. This is the hypergeometric setting; when the sample is small relative to the deck, the binomial is a reasonable approximation. (c) **Yes (approximately)** — $n = 100$ fixed, two outcomes, and p is essentially constant because 100 is a small fraction of the adult population (the 10% rule for independence). (d) **No** — n is *not fixed*; the number of trials is itself random. This is the geometric distribution. (e) **Yes (approximately)** — the four conditions hold as long as bulbs are independent (no shared manufacturing defect that would make failures cluster).

(a) At $n = 20$, $p = 0.05$ the PMF is strongly right-skewed; most mass sits at 0–3 successes. The normal overlay is a poor fit — it extends to negative values and is symmetric, while the binomial cannot go below 0 and is asymmetric. (b) At $n = 200$, $p = 0.05$ the PMF is approximately bell-shaped and centered near 10; the normal overlay fits well. (c) Case (a): $np = 1$ fails the rule. Case (b): $np = 10$ and $n(1-p) = 190$ — both pass. (d) The rule needs *both* pieces because a distribution can be skewed either way. At $p = 0.99$, $n = 100$: $np = 99$ passes easily, but $n(1-p) = 1$ fails — the distribution is squished against its upper bound and is left-skewed. Checking only np would miss this.

(a) Expected wins $= np = 45$. Probability of coming out ahead: $P(X \geq 51) = 1 - P(X \leq 50) \approx 0.134$, about 13%. (b) The law of large numbers says your average per-round result will converge to the expected value: $E[X] - n/2 = 45 - 50 = -\$5$ per round (equivalently $-\$0.05$ per trial). Random "good" rounds are cancelled out by random "bad" rounds in the long run. (c) The expected loss is small *per round*, but with independent rounds the LLN guarantees that total losses grow linearly with the number of rounds — exactly the opposite of what the player wants to believe. The "small per round" framing hides the fact that the house edge is paid *every* round. (d) The probability of coming out ahead **shrinks** as n grows: about 13% at $n = 100$, well under 1% at $n = 1000$, essentially 0 at $n = 10{,}000$. As n grows, $\hat{p}$ concentrates around the true $p = 0.45$, so $\hat{p} > 0.5$ becomes vanishingly rare.

E.28 Chapter 28 — More About the Normal Distribution

Solutions to the odd-numbered exercises in Chapter 28 — More About the Normal Distribution.

(a) **Approximately normal** — a straight-line QQ plot is exactly what we hope to see; the observed quantiles match the normal quantiles at every point. (b) **Right-skewed** — the upper tail extends farther than a normal would predict, so the largest observed values exceed the normal quantiles and the points rise above the line. (c) **Left-skewed** — the lower tail extends farther than normal, so the smallest observed values fall below the normal quantiles and the points drop below the line on the left. (d) **Heavy-tailed** — both extremes are more extreme than a normal would predict (typical of a t-distribution with low df, or data with occasional outliers on both sides).

(a) **Consistent with normality.** With $n = 15$ we *expect* some scatter around the line even when the data truly are normal — random sampling variability alone produces wobble at small n. The signal that matters is the presence or absence of a systematic curve (S-shape, monotonic bend at one end), not the amount of wobble. (b) **Yes, the conclusion changes.** With $n = 500$ random sampling variability is much smaller and the plot should look much cleaner; persistent scatter at that sample size would indicate a real departure from normality (mixture distribution, heteroscedasticity, etc.). (c) **No — the reasoning is unsound.** The t-test is robust to moderate departures from normality, especially with modest-to-large n; perfect linearity of the QQ plot is not required. The right questions are whether the departure is severe (clear skew or heavy tails) and whether n is small enough for departures to matter. Modest wobble at any n is not a reason to abandon the t-test.

(a) **Agree.** Both diagnostics point to strong right skew. The histogram makes the shape visible directly; the QQ plot's upward curve at the right end confirms an over-long upper tail relative to a normal. (b) A **log transformation** ($\log(x + \varepsilon)$, or $\sqrt{x}$ for zero-heavy data) is a natural choice for right-skewed positive data — it compresses the long upper tail and expands the crowded region near zero, often producing an approximately symmetric transformed distribution. (c) **Cautiously trust the CI, with the caveat that "the mean" may not be the interesting quantity.** With $n = 300$ the CLT makes the sampling distribution of $\bar{x}$ approximately normal even from skewed data, so the t CI's coverage is roughly correct. The deeper caveat: for very skewed data the mean is often not the best summary of a typical day — the median or another percentile may better answer the substantive question.

StatLens Exercises

(a) About **0.683** (68.3%) — the "68%" rule is a rounded version of the exact area. (b) About **0.954** (95.4%) — again, the rule rounds; to get exactly 95% you would need ± 1.96 SDs, not ± 2. (c) About **0.997** (99.7%) — essentially exact here. (d) The **curve is the truth**; the empirical rule is a memorable approximation. When precision matters (a CI at exactly 95%, for instance) use the curve — or the tool — directly, not the ± 2 shortcut.

(a) Right-tail area = 0.10 → $z \approx 1.28$ → score $= 500 + 1.28 \cdot 100 = 628$. (b) Left-tail area = 0.25 → $z \approx -0.67$ → score $= 500 - 67 = 433$. (c) 75th percentile $\approx 500 + 0.67 \cdot 100 = 567$, so IQR $= 567 - 433 = 134$ points. (d) **Not surprising.** The 90th percentile is at 628 (from part a). A 50-point band is half a standard deviation, and near the cutoff the normal density is still high, so a 50-point window can contain many rejected applicants. With $\sigma = 100$, "near the cutoff" really is a wide window.

(a) Height is roughly the **sum of many small genetic and developmental contributions**, each acting independently. The Central Limit Theorem for sums of many small independent effects delivers approximate normality. (b) Hourly wages are strongly **right-skewed** — most workers earn modest hourly rates, with a long upper tail from high earners (surgeons, executives) pulling the mean well above the median. Expect a shape that peaks low and trails off far to the right. (c) A normal model would **under-predict the probability of very high wages** (the normal's symmetric tails thin out too quickly) and **over-predict wages near the mean** (a symmetric peak where the actual distribution is asymmetric). (d) The normal still describes the **sampling distribution of** $\bar{x}$ by the CLT, even when the population is skewed — provided n is large enough. You can build a normal-based CI for the *mean* wage even though individual wages are not normal. Inference about individual wages (e.g., "what wage is at the 95th percentile?") requires a different model.

E.29 Chapter 29 — Type II Error and Power

Solutions to the odd-numbered exercises in Chapter 29 — Type II Error and Power.

(a) Scenario (i) is higher. A smaller sample size gives a less precise estimate of $\hat{p}$, so its standard error is larger. (b) Scenario (i) is higher. Increasing the confidence level widens the interval, which requires a larger margin of error. (c) They are equal. Once the Z-statistic is fixed, the p-value depends only on the tail area beyond it — the sample size is already baked into the Z-statistic itself. (d) Scenario (i) is higher. Lowering α makes H_0 harder to reject, so when H_A is truly in force we are more likely to fail to detect it (a larger Type II error rate, β).

True. As the sample size grows the standard error shrinks toward zero. With a small enough standard error, even a tiny gap between the observed point estimate and the null value produces a large Z-statistic and a small p-value. Statistical discernibility is not the same as practical importance — a discernible difference can still be too small to matter.

StatLens Exercises

(a) β is shaded under the **alternative** curve, in the region *below* the critical value (the "do not reject H_0" region). It lives under H_A because β is defined as the probability of failing to reject when the alternative is true. (b) "$\beta = 0.14$" means that if the alternative is really in force at this effect size, this experiment has a 14% chance of *missing* the effect — a 14% chance of concluding 'no evidence' when there truly is an effect. (c) Decreasing δ to 0.5 makes β **rise** sharply. A smaller effect pushes the alternative curve closer to the null, so much more of the alternative distribution falls into the do-not-reject region. (d) Decreasing σ to 1 makes β **fall**. Less noise sharpens both curves and pulls their peaks apart, so less of the alternative distribution overlaps the do-not-reject region.

(a) At $\delta = 0.3,\ n = 30,\ \sigma = 1,\ \alpha = 0.05$ the power is roughly 0.50 — essentially a coin flip. A trial with this little power is *not* informative about whether the effect is real: "no detected effect" is exactly what we'd expect half the time even if the treatment truly worked at $\delta = 0.3$. (Recall the ≥ 0.80 rule of thumb for adequate power.) (b) At $\delta = 1$ the power is essentially 1.0. A non-result at this level of power *is* informative — it rules out effects that large with high confidence. (c) "**No evidence of effect**" means the data did not push us to reject H_0; this can happen because there truly is no effect *or* because the study lacked power to detect it. "**Evidence of no effect**" is a positive claim that any effect must be small, and requires enough power to have detected a meaningful effect if it existed. The two are not interchangeable. (d) Something like: "We failed to detect an effect of $\delta \geq 1$ (power ≥ 0.95 against effects that large), but were underpowered to detect smaller effects (power ≈ 0.50 against $\delta = 0.3$). A replication with $n \geq 200$ would be needed to draw firm conclusions about smaller, clinically meaningful effects."

E.30 Chapter 30 — Statistical Power

Solutions to the odd-numbered exercises in Chapter 30 — Statistical Power.

(a) (i) is larger. $SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$ shrinks as n grows, so the smaller $n = 125$ gives the larger standard error. (b) (i) is larger. A higher confidence level requires a larger $z^\star$, which widens the margin of error. (c) Equal. The p-value depends only on the Z-statistic and the direction of the test — not on the sample size that produced it. (d) (ii) is larger. Increasing the discernibility level (α) enlarges the rejection region under H_0, which shrinks the rejection region under H_A and therefore raises β.

True. As n grows, the standard error of the point estimate shrinks like $1/\sqrt{n}$, so the test statistic grows for a fixed observed difference and the p-value falls. With a large enough sample, a difference that is far too small to matter in context can still be flagged as statistically discernible — which is why practical importance (effect size, context) must be judged separately from a p-value.

StatLens Exercises

(a) The left curve is the null distribution (centered at $\mu = 0$); the right curve is the alternative (centered at $\mu = \delta = 1$). Each curve is centered at the mean it assumes. (b) α is the right-tail area of the null curve past the critical value (the false-alarm region under H_0). β is the area of the alternative curve to the *left* of the critical value (the missed-detection region under H_A). Power $= 1 - \beta$ is the area of the alternative curve to the *right* of the critical value. (c) At $\alpha = 0.05$, $n = 30$, $\delta = 1$, $\sigma = 2$, the tool reports power ≈ 0.756. (d) "If a real effect of this size exists, this experiment has only a 50/50 chance of detecting it" — roughly a coin flip's worth of detection ability.

(a) With 500 simulated tests, the empirical reject rate is close to the analytic power of ≈ 0.76, typically within a few percentage points of Monte Carlo error. (b) Yes. Monte Carlo error scales like $1/\sqrt{N}$, so 5000 sims tighten the estimate by roughly a factor of 3 compared to 500. (c) Under a real effect, p-values cluster near 0. Under H_0 they are uniform on $[0, 1]$; under H_A the distribution shifts toward small values, and the cluster gets denser as the effect size or n grows. (d) (i) Trust calibration: the analytic formula assumes normality, equal variances, and independence; simulating with the actual generating process tells you whether the analytic power is honest when those assumptions bend. (ii) Intuition building: watching 500 individual studies that "ought to detect a real effect" but about 25% still miss it teaches the meaning of power more viscerally than reading "power $= 0.76$" off a tool.

(a) About 80 labs report a statistically discernible result (power $= 0.8$ means an 80% rejection rate when H_A is true), and about 20 report a non-result. (b) Publication bias. If only the 80 "discernible" labs publish, the literature is enriched for positive findings, and a reader who sees only those 80 papers overestimates the strength of the effect. Meta-analyses that draw only on published studies inherit the same bias. (c) A meta-analysis of all 100 labs sees the true effect with a much smaller standard error (pooled n is 100 times larger) and can also detect publication bias by comparing the shape of published-vs-unpublished results. Single-study significance is fragile; meta-analysis is the right level of inference. (d) A power-0.5 study is a coin flip for detection *if the effect exists* — and a headline conflates "detected" with "exists." Given publication bias and the file-drawer problem, a single low-power "significant" result is weak evidence; wait for replication before believing the claim.

E.31 Chapter 31 — Sample Size Planning

Solutions to the odd-numbered exercises in Chapter 31 — Sample Size Planning.

StatLens Exercises

(a) Slide n upward on the Power & Error Visualizer until the power readout crosses 0.80. This happens at about $n \approx 40$ (the exact value depends on the tool's tail conventions). (b) About $n \approx 55$ achieves 90% power. Each additional percentage point of power costs more n than the last, because power approaches its ceiling of 1.0 asymptotically — the "last 10%" is always more expensive than any earlier 10%. (c) Much smaller — about $n \approx 10$. Larger true effects are easier to detect, so a bigger δ/σ dramatically reduces the n required for the same

power. (d) The required n depends on the *true* effect size, which is unknown before the study. With no prior estimate, a researcher can: (i) use the **smallest effect of practical or clinical interest**, so the study is powered to detect anything worth caring about; (ii) run a **pilot study** first to estimate σ and a plausible δ; or (iii) fix n and instead report the **minimum detectable effect** at the planned n.

(a) $n_{\text{per group}} = 2 \cdot (2.8/0.5)^2 = 2 \cdot (5.6)^2 = 62.7 \to 63$ per group. (b) The Power & Error Visualizer models a *one-sample* test, so its slider crosses 0.80 near $n \approx 32$ — roughly *half* of 63. That is not a contradiction: a two-sample comparison needs about **twice** the one-sample n because the two-sample standard error carries a $\sqrt{2}$ factor. The tool confirms the *logic* (power rises past 0.80), but read its n as the one-sample equivalent, not the per-group count for two samples. (c) Divide by $1 - 0.20 = 0.80$: enroll $63/0.80 \approx 79$ per group (round up to be safe). (d) At $n = 30$, $\delta/\sigma = 0.3$, $\alpha = 0.05$ two-sided, power is only about 0.21 — a 21% chance of detecting a real small effect. **Do not trust** a non-result from this study; it was severely underpowered.

Adolph, S. C. 1987. "Physiological and Behavioral Ecology of the Lizards *Sceloporus Occidentalis* and *Sceloporus Graciosus*." PhD thesis, University of Washington, Seattle, Washington.

———. 1990. "Influence of Behavioral Thermoregulation on Microhabitat Use by Two *Sceloporus* Lizards." *Ecology* 71: 315–27.

Allais, G., M. Romoli, S. Rolando, G. Airola, I. Castagnoli Gabellari, R. Allais, and C. Benedetto. 2011. "Ear Acupuncture in the Treatment of Migraine Attacks: A Randomized Trial on the Efficacy of Appropriate Versus Inappropriate Acupoints." *Neurological Sciences* 32 (1): 173–75. https://doi.org/10.1007/s10072-011-0525-4.

Allison, T., and D. V. Cicchetti. 1975. "Sleep in Mammals: Ecological and Constitutional Correlates." *Arch. Hydrobiol* 75: 442. https://doi.org/10.1126/science.982039.

American Council on Education. 2008. "College-Bound Students' Interests in Study Abroad and Other International Learning Activities." http://www.openintro.org/redirect.php?go=textbook-Interests_in_Study_Abroad_2008.

Asbury, D. A., and S. C. Adolph. 2007. "Behavioral Plasticity in an Ecological Generalist: Microhabitat Use by Western Fence Lizards." *Evolutionary Ecology Research* 9: 801–15.

Audera, C., R. V. Patulny, Sander B. H, and R. M. Douglas. 2001. "Mega-Dose Vitamin c in Treatment of the Common Cold: A Randomised Controlled Trial." *Medical Journal of Australia* 175 (7): 359–62. http://www.openintro.org/redirect.php?go=textbook-vitamin_C_cold_treatment_2001.

Backstrom, L. 2011. "Anatomy of Facebook." *Facebook Data Team's Notes*. http://www.openintro.org/redirect.php?go=textbook-anatomy-of-facebook.

Benson, J. B. 1993. "Season of Birth and Onset of Locomotion: Theoretical and Methodological Implications." *Infant Behavior and Development* 16 (1): 69–81. https://doi.org/10.1016/0163-6383(93)80029-8.

Bucciol, Alessandro, and Marco Piovesan. 2011. "Luck or Cheating? A Field Experiment on Honesty with Children." *Journal of Economic Psychology* 32 (1): 73–78. https://doi.org/10.1016/j.joep.2010.12.001.

CDC. 2008. "Perceived Insufficient Rest or Sleep Among Adults – United States, 2008." http://www.openintro.org/redirect.php?go=textbook-Perceived_Insufficient_Rest_or_Sleep_Among_Adults.

———. 2018. "2018 Assisted Reproductive Technology Fertility Clinic Success Rates Report." https://www.cdc.gov/art/pdf/2018-report/ART-2018-Clinic-Report-Full.pdf.

Conner, T. S., K. L. Brookie, A. C. Carr, L. A. Mainvil, and M. CM. Vissers. 2017. "Let Them Eat Fruit! The Effect of Fruit and Vegetable Consumption on Psychological Well-Being in Young Adults: A Randomized Controlled Trial." *PloS One* 12 (2). https://doi.org/10.1371/journal.pone.0171206.

Datoo, M. S., M. H. Natama, A. Somé, O. Traoré, T. Rouamba, D. Bellamy, P. Yameogo, et al. 2021. "High Efficacy of a Low Dose Candidate Malaria Vaccine, R21 in 1 Adjuvant Matrix-m^{TM}, with Seasonal Administration to Children in Burkina Faso." *The Lancet*. https://doi.org/10.1016/S0140-6736(21)00943-0.

Demos. 2011. "The State of Young America: The Poll." http://www.openintro.org/redirect.php?go=textbook-young_americans_2011_extra.

Frederick, S., N. Novemsky, J. Wang, R. Dhar, and S. Nowlis. 2009. "Opportunity Cost Neglect." *Journal of Consumer Research* 36 (4): 553–61.

Gallup. 2012. http://www.openintro.org/redirect.php?go=textbook-employed_americans_in_better_health_2012.

Gallup. 2021a. "Half of College Students Say COVID-19 May Impact Completion." https://www.openintro.org/go?id=textbook-gallup-2021-covid-college-impact.

———. 2021b. "U.s. Support for Vaccination Proof Varies by Activity, Data Collected in April 2021." https://www.openintro.org/go?id=textbook-gallup-2021-vaccine-proof.

Garbutt, J. M., C. Banister, E. Spitznagel, and J. F. Piccirillo. 2012. "Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial." *JAMA: The Journal of the American Medical Association* 307 (7): 685–92. https://doi.org/10.1001/jama.2012.138.

Gneezy, Uri, and Aldo Rustichini. 2000. "A Fine Is a Price." *The Journal of Legal Studies* 29 (1): 1–17. https://doi.org/10.1086/468061.

Gorman, K. B., T. D. Williams, and W. R. Fraser. 2014. "Ecological Sexual Dimorphism and Environmental Variability Within a Community of Antarctic Penguins (Genus Pygoscelis)." *PLoS ONE* 9(3) (e90081): –13. https://doi.org/10.1371/journal.pone.0090081.

Graham, D. J., R. Ouellet-Hellstrom, T. E. MaCurdy, F. Ali, C. Sholley, C. Worrall, and J. A. Kelman. 2010. "Risk of Acute Myocardial Infarction, Stroke, Heart Failure, and Death in Elderly Medicare Patients Treated with Rosiglitazone or Pioglitazone." *JAMA* 304 (4): 411. https://doi.org/10.1001/jama.2010.920.

Heinz, G., L. J. Peterson, R. W. Johnson, and C. J. Kerk. 2003. "Exploring Relationships in Body Dimensions." *Journal of Statistics Education* 11 (2). http://www.openintro.org/redirect.php?go=textbook-body_dim_2003.

Hepler, J., and D. Albarracín. 2013. "Attitudes Without Objects: Evidence for a Dispositional Attitude, Its Measurement, and Its Consequences." *Journal of Personality and Social Psychology* 104 (6): 1060. https://doi.org/10.1037/a0032282.

ICPSR. 2014. "United States Department of Health and Human Services. Centers for Disease Control and Prevention. National Center for Health Statistics. Natality Detail File, 2014 United States. Inter-University Consortium for Political and Social Research, 2016-10-07." https://doi.org/10.3886/ICPSR36461.v1.

Kaiser Family Foundation. 2019. https://www.openintro.org/go?id=textbook-kff-2019-expand-aca-coverage.

King, C., J. Suamani, N. Sanuku, Y.-C. Cheng, S. Satofan, B. Mancuso, C. W. Goss, et al. 2018. "A Trial of a Triple-Drug Treatment for Lymphatic Filariasis." *New England Journal of Medicine* 379: 1801–10. https://doi.org/10.1056/NEJMoa1706854.

Latter, O. H. 1902. "The Egg of Cuculus Canorus. An Enquiry into the Dimensions of the Cuckoo's Egg and the Relation of the Variations to the Size of the Eggs of the Foster-Parent, with Notes on Coloration." *Biometrika* 1: 164–76. https://doi.org/10.2307/2331486.

Legaki, Nikoletta-Zampeta, Nannan Xi, Juho Hamari, Kostas Karpouzis, and Vassilios Assimakopoulos. 2020. "The Effect of Challenge-Based Gamification on Learning: An Experiment in the Context of Statistics Education." *International Journal of Human-Computer Studies* 144. https://doi.org/10.1016/j.ijhcs.2020.102496.

Lo Sasso, A. T., M. R. Richards, C. F. Chou, and S. E. Gerber. 2011. "The $16,819 Pay Gap for Newly Trained Physicians: The Unexplained Trend of Men Earning More Than Women." *Health Affairs* 30 (2). https://doi.org/10.1377/hlthaff.2010.0597.

Lockman, S., R. L. Shapiro, L. M. Smeaton, C. Wester, I. Thior, L. Stevens, F. Chand, et al. 2007. "Response to Antiretroviral Therapy After a Single, Peripartum Dose of Nevirapine." *New England Journal of Medicine* 356 (2): 135–47. https://doi.org/10.1056/NEJMoa062876.

Lucas, M., F. Mirzaei, A. Pan, O. I. Okereke, W. C. Willett, E. J. O'Reilly, K. Koenen, and A. Ascherio. 2011. "Coffee, Caffeine, and Risk of Depression Among Women." *Archives of Internal Medicine* 171 (17): 1571. https://doi.org/10.1001/archinternmed.2011.393.

Malkevitch, J., and L. M. Lesser. 2008. *For All Practical Purposes: Mathematical Literacy in Today's World*. WH Freeman & Co.

McNeil, D. R. 1977. "Interactive Data Analysis: A Practical Primer."

Mortada, WI, MA Sobh, MM El-Defrawy, and SE Farahat. 2000. "Study of Lead Exposure from Automobile Exhaust as a Risk for Nephrotoxicity Among Traffic Policemen." *American Journal of Nephrology* 21 (4): 274–79. https://doi.org/10.1159/000046261.

National Sleep Foundation. 2012. http://www.openintro.org/redirect.php?go=textbook-trans_workers_sleep_2012.

Nieman, D. C., E. J. Cayea, M. D. Austin, D. A. Henson, S. R. McAnulty, and F. Jin. 2009. "Chia Seed Does Not Promote Weight Loss or Alter Disease Risk Factors in Overweight Adults." *Nutrition Research* 29 (6): 414–18. https://doi.org/10.1016/j.nutres.2009.05.011.

Nightingale, Florence. 1859. *Notes on Nursing: What It Is, and What It Is Not*. Harrison & Sons.

NOAA. 2023. "Climate Data Online." http://www.openintro.org/redirect.php?go=textbook-noaa_1948_2018.

NORC. 2010. http://www.openintro.org/redirect.php?go=textbook-gss_2010.

———. 2016. "Data Explorer - Government Resp to Promote Gender Equality." https://www.openintro.org/go?id=textbook-gssdataexplorer-marijuana-gender-equality.

———. 2022. "Data Explorer - Should Marijuana Be Made Legal." https://www.openintro.org/go?id=textbook-gssdataexplorer-marijuana-legal.

O'Connor, A. 2011. "Coffee Drinking Linked to Less Depression in Women." *New York Times*. http://www.openintro.org/redirect.php?go=textbook-coffee_depression_2011.

Oldham-Cooper, R. E., C. A. Hardman, C. E. Nicoll, P. J. Rogers, and J. M. Brunstrom. 2011. "Playing a Computer Game During Lunch Affects Fullness, Memory for Lunch, and Later Snack Intake." *The American Journal of Clinical Nutrition* 93 (2): 308. https://doi.org/10.3945/ajcn.110.004580.

Orben, A., and A. K. Baukney-Przybylski. 2018. "Screens, Teens and Psychological Well-Being: Evidence from Three Time-Use Diary Studies." *Psychological Science*. https://doi.org/10.1177/0956797619830329.

Othman, S., N. Hassan, and A. Mohamed. 2023. "Effectiveness of Mindfulness-Based Interventions on Burnout and Self-Compassion Among Critical Care Nurses Caring for Patients with COVID-19: A Quasi-Experimental Study." *BMC Nursing* 22: 35.

Parker-Pope, T. 2011. "The School Bully Is Sleepy." *New York Times*. http://www.openintro.org/redirect.php?go=textbook-school_bully_sleepy_2011.

Pew Research Center. 2011. "Is College Worth It?, Data Collected March 15-29, 2011." http://www.openintro.org/redirect.php?go=textbook-college_worth_it_2011.

———. 2013. "The Diagnosis Difference." https://www.openintro.org/go?id=textbook-pew-2013-diagnosis-difference.

———. 2018. "A Majority of Teens Have Experienced Some Form of Cyberbullying, Data Collected March 7 - April 10, 2018." https://www.openintro.org/go?id=textbook-pew-2018-cyberbullying.

———. 2021a. "How Americans' Attitudes about Climate Change Differ by Generation, Party and Other Factors, Data Collected April 20 to 29, 2021." https://www.openintro.org/go?id=textbook-pew-2021-climate-change.

———. 2021b. "The State of Online Harassment, Data Collected September 8 - 13, 2020." https://www.openintro.org/go?id=textbook-pew-2021-online-harassment.

———. 2022. "News Consumption Across Digital Platforms." https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/pj_2022-09-20_news-platforms_0-03b/.

Pfizer. 2021. "Pfizer-Biontech Announce Positive Topline Results of Pivotal Covid-19 Vaccine Study in Adolescents." https://www.openintro.org/go?id=textbook-pfizer-2021-adolescent-covid-vaccine.

Pierson, Emma, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, et al. 2020. "A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States." *Nature Human Behaviour* 4 (7): 736–45. https://doi.org/10.1038/s41562-020-0858-1.

Piff, P. K., D. M. Stancato, S. Côté, R. Mendoza-Denton, and D. Keltner. 2012. "Higher Social Class Predicts Increased Unethical Behavior." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1118373109.

Poll, Marist. 2011. "Road Rules: Re-Testing Drivers at Age 65?, March 4, 2011." http://www.openintro.org/redirect.php?go=textbook-drivers_at_65_2011.

Polling, Public Policy. 2015. "Americans on College Degrees, Classic Literature, the Seasons, and More, Data Collected Feb 20-22, 2015." http://www.openintro.org/redirect.php?go=textbook-PPP_30215.

Rabin, R. C. 2010. "Risks: Smokers Found More Prone to Dementia." *New York Times*. http://www.openintro.org/redirect.php?go=textbook-nytimes_smoking_dementia.

Rao, M., L. Bar, Y. Yu, M. Srinivasan, A. Mukherjea, J. Li, S. Chung, S. Venkatraman, S. Dan, and L. Palaniappan. 2021. "Disaggregating Asian American Cigarette and Alternative Tobacco Product Use: Results from the National Health Interview Survey (NHIS) 2006–2018." *Journal of Racial and Ethnic Health Disparities*. https://doi.org/10.1007/s40615-021-01024-5.

Ritz, B., F. Yu, G. Chapa, and S. Fruin. 2000. "Effect of Air Pollution on Preterm Birth Among Children Born in Southern California Between 1989 and 1993." *Epidemiology* 11 (5): 502–11. http://www.jstor.org/stable/3703990.

Rosen, B., and T. H. Jerdee. 1974. "Influence of Sex Role Stereotypes on Personnel Decisions." *Journal of Applied Psychology* 59 (1): 9. https://doi.org/doi.org/10.1037/h0035834.

Scanlon, T. J., R. N. Luben, F. L. Scanlon, and N. Singleton. 1993. "Is Friday the 13th Bad for Your Health?" *British Medical Journal* 307: 1584–86. https://doi.org/10.1136/bmj.307.6919.1584.

Smith, G. E., P. A. Chouinard, and S. Byosiere. 2021. "If i Fits i Sits: A Citizen Science Investigation into Illusory Contour Susceptibility in Domestic Cats (*Felis Silvestris Catus*)." *Applied Animal Behaviour Science* 240. https://doi.org/10.1016/j.applanim.2021.105338.

Survey USA. 2012. "News Poll 19333, Data Collected on June 27, 2012." http://www.openintro.org/redirect.php?go=textbook-SurveyUSA_19333.

———. 2019. http://www.openintro.org/redirect.php?go=textbook-SurveyUSA_24568.

———. 2021. https://www.openintro.org/go?id=textbook-surveyusa-2021-seattle-defund-police.

Turnbull, B., B. Brown, and M. Hu. 1974. "Survivorship of Heart Transplant Data." *Journal of the American Statistical Association* 69: 74–80. https://doi.org/10.1080/01621459.1974.10480130.

UNC Carolina Population Center. 2006. "China Health and Nutrition Survey, 2006." http://www.openintro.org/redirect.php?go=textbook-china_health_nut_survey_2006.

US Census Bureau. 2010. "2010 Census Urban and Rural Classification and Urban Area Criteria Housing Characteristics." http://www.openintro.org/redirect.php?go=textbook-housing_char_2010.

US DOE EPA. 2021. "Fuel Economy, 2021 Data File." https://www.openintro.org/go?id=textbook-fueleconomy-2021.

Vaughn, A. 2011. "Poll Finds Young Adults Optimistic, but Not about Money." *Los Angeles Times*. http://www.openintro.org/redirect.php?go=textbook-young_americans_2011.

Washington Post. 2020. "Washington Post-Schar School National Poll, Data Collected June 2-7, 2020." http://www.openintro.org/redirect.php?go=exercise-wapo-2020-blm-poll.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

You, Wenpeng, Renata Henneberg, Arthur Saniotis, Yanfei Ge, and Maciej Henneberg. 2022. "Total Meat Intake Is Associated with Life Expectancy: A Cross-Sectional Data Analysis of 175 Contemporary Populations." *International Journal of General Medicine* 15: 1833–51. https://doi.org/10.2147/IJGM.S333004.

YouGov. 2021. "Do You Think the Coronavirus Will End up Bringing the World Closer Together, or Leave Us Further Apart?" https://www.openintro.org/go?id=textbook-yougov-2021-covid.